AD_____

Award Number DAMD17-99-1-9318

TITLE: Rational structure-based design of anti-breast-cancer drugs targeting the erbB family of receptor tyrosine

PRINCIPAL INVESTIGATOR: Ruben Abagyan, Ph.D.

CONTRACTING ORGANIZATION: The Scripps Research Institute
La Jolla, California 92037

REPORT DATE: September 2000

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2000	Annual (1 Sep 99 - 31 Aug 00)

4. TITLE AND SUBTITLE

Rational structure-based design of anti-breast-cancer drugs targeting the erbB family of receptor tyrosine kinases

5. FUNDING NUMBERS

DAMD17-99-1-9318

6. AUTHOR(S)

Ruben Abagyan, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Scripps Research Institute
La Jolla, California 92037

E-Mail: Abagyan@ Scripps.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Report contains color photos

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

The new inhibitors of the ErbB2 tyrosine kinase domain are potential lead compounds against the most aggressive forms of breast cancer. To inhibit the catalytic domain of the receptor, we intended to crystallize the tyrosine kinase domain of the receptor and identify new lead candidates through a new virtual ligand screening procedure. The key component and prerequisite of this technology is the three-dimensional model of the target domain. This structure can either be obtained through an X-ray crystallography experiment, the best case scenario, or through model building by homology to other know structures. The next step is to improve the flexible docking procedure and perform it in a high throughput manner to predict the binders of the target of interest. After preliminary work on microcrystallization we expect a reasonable chance to crystallize the ErbB2 domain, even though it remains extremely challenging. Using the above technology, we are now ready to begin crystallization trials. We are ready to start docking studies using the models of ErbB2. These studies should first explain the pattern of cross-reactivity with other tyrosine kinase ligands. In the event that we obtain a crystal structure we will be ready to use it immediately for virtual ligand screening.

14. SUBJECT TERMS

breast cancer, computer modeling, x-ray crystallography, receptor Tyrosine kinases, rational drug design

15. NUMBER OF PAGES

73

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

Table of Contents

INTRODUCTION

The new inhibitors of the ErbB2 tyrosine kinase domain are potential lead compounds against the most aggressive forms of breast cancer. To inhibit the catalytic domain of the receptor, we intended to crystallize the tyrosine kinase domain of the receptor and identify new lead candidates through a new virtual ligand screening procedure. The key component and prerequisite of this technology is the three-dimensional model of the target domain. This structure can either be obtained through an X-ray crystallography experiment, the best case scenario, or through model building by homology to other know structures. The next step is to improve the flexible docking procedure and perform it in a high throughput manner to predict the binders of the target of interest. We have pursued all three directions: tried to crystallize the domain, built several models of ErbB2 and worked on the optimization of parameters for the flexible docking calculations.

BODY

Task 1. To determine a high-resolution crystal structure of the tyrosine kinase domain of ErbB2 (ErbB2-TKD).

The experimental determination of the ErbB2-TKD has been pursued by many laboratories and companies, but so far no one has reported a success. The main bottleneck is obtaining a reasonably large diffracting crystal. In that effort we focused on an attempt to take advantage of the new microstallization technology which allows one to try many different crystal conditions at a time.

We have spent the past year developing high throughput technologies to aid in the structure determination of ErbB2. These technologies have been designed to allow for high throughput co-crystal structure determinations, specifically to allow one to determine a large number of inhibitor compounds that are predicted based on the virtual ligand screening. Once a series of compounds have been found to be potential inhibitors by virtual ligand screening and are confirmed by cell-based assays, we will be able to quickly place these compounds into the robotics system and accurately determine their binding site.

Using technology developed at Lawrence Berkeley National Laboratory and relocated to the Genomics Institute for the Novartis Research Foundation last year, we have determined that we can reproducibly create 20 nanoliter droplets for protein crystallization trials. This current system evaluates a screen of 480 different crystallization conditions (thus requiring 9.6 microliters per screen at a typical protein concentration of 10 mgs/ml). In addition, we have developed an imaging system capable of imaging and analyzing up to 138,000 trials per day. Since ErbB2 is expressed in low quantities in a baculovirus expression system, this nanoliter volume technology for protein crystallization will be extremely useful for the crystallization of the protein target.

We have also designed a novel plate that uses the nanoliter volumes described above and is capable of being placed into a high throughput screening system and collecting compounds in the same location as the crystallization trial. This will allow one to quickly screen through potential inhibitors in a fast and efficient manner. The plates use a standard 96 well format, but have a unique design to allow them to both collect screened compounds but also conduct crystallization experiments.

Task 2. To build structural models by homology for the kinase domains

In the absence of the experimental three-dimensional structure of the ErbB2 TKD, we embarked on the homology study of this domain. There are several different kinase domains with known three-dimensional structure, but all of them fall in the weak zone of sequence similarity between 30 and 40% (Fig. 1). The domains include: the FGFR TK (fibroblast growth factor receptor), the insulin receptor and its mutant, the LCK kinase, SRC tyrosine kinase and haematopeotic cell kinase HCK (see Fig. 1)

To estimate the error in the model we built the alignments and models starting from two different templates, the fibroblast growth factor receptor represented by the 1fgk entry in the protein databank and the SRC tyrosine kinase 1fmk entry. The first model was built on the basis of the alignment presented in Figure 2. In this alignment we built all the missing loops and where necessary the loops have been expanded and searched against an entire pdb database.

Figure 3 compares the template and the model and shows the amino acids that are different between the two.

The analysis of the active site alone shows that the SRC kinase is a slightly better template than the FGFR tyrosine kinase domain. The local residue conservation in the active site is higher for the SRC template. We have built another model of ErbB2 to be able to compare the two models built using different three-dimensional templates.

Figure 4 presents a binding site of the model with docked ATP analogue. The binding site of ErbB2 is distinctly different from both the SRC kinase binding site (Figure 5) and the FGFR binding site. Both in SRC and FGFR domains tyrosine 340 (the number is taken from 2src structure) is replaced by leucine in ErbB2 domain. Another essential difference: alanine 403 in both template structures is replaced by threonine, and V561 of the FGFR structure is replaced by threonine.

The differences we observed as a result of modeling can be used to design molecules specific to ErbB2 domain as opposed to the other kinases. The threonine substitutions can be specifically targeted by a number of hydrogen bond donors and acceptors of a lead compound. The leucine substitution leads to a large hydrophobic pocket that can also be targeted in our design. We are currently working on better ways to exploit small differences in the active sites.

Improvements of the docking technology.

We have used docking technology from Molsoft. Several attempts have been made to estimate the accuracy of the procedure and validate the technology, which we intended to use to search for new lead compounds. The best validation of the technology was a real application of flexible docking to the design of RAR antagonists (Ref. 1) and agonists (Ref 2.).

In manuscripts Ref. 3 and Ref. 4 we have demonstrated how the parameters of the docking and scoring functions can be optimized using the known protein-ligand complexes. In a separate effort we collaborated with the laboratory of Prof. Charles Brooks III, who compared four different docking methods which could have been used for the virtual ligand screening step of this project. Our original approach, using ICM, has proved to be the best docking and screening procedure.

KEY RESEARCH ACCOMPLISHMENTS

1) The microcrystallization robotic system has been tested and optimized. The nanoliter crystallization will allow to try many different crystallization conditions for ErbB2
2) The models were built for ErbB2 based on two different templates with known three-dimensional structures.
3) The binding site has been compared between the ErbB2 and FGFR tyrosine kinases.
4) The ICM docking procedure has been compared with other docking algorithms and is ready to be used for lead discovery.

REPORTABLE OUTCOMES

A web site with ErbB2 models has been created.
The manuscript comparing different flexible docking methods and the virtual ligand screening algorithm has been prepared for publication (Ref. 5).

CONCLUSIONS

After preliminary work on microcrystallization we expect a reasonable chance to crystallize the ErbB2 domain, even though it remains extremely challenging. Using the above technology, we are now ready to begin crystallization trials.

We are ready to start docking studies using the models of ErbB2. These studies should first explain the pattern of cross-reactivity with other tyrosine kinase ligands.

In the event that we obtain a crystal structure we will be ready to use it immediately for virtual ligand screening.

REFERENCES

1. Schapira, M., Raaka, B.M., Samuels, H, H. and Abagyan, R. (2000). Rational design of novel nuclear hormone receptor antagonists. *PNAS* **97** (3), 1008-1013.

2. Schapira, M., Raaka, B. M., Samuels, H.H., and Abagyan, R. (2000). In silico discovery of novel RAR agonist structures. *J. Med. Chem.* (submitted).

3. Totrov, M., and Abagyan R. (1999). Derivation of sensitive discrimination potential for virtual ligand screening. *Proceedings of the Third Annual Intl. Conf. on Comp. Mol. Bio.* 312-320.

4. Totrov, M., and Abagyan R. (2000). Protein ligand docking as an energy optimization problem. Thermodynamics of the drug receptor interactions.*(R.B. Raffa, ed.),(in press).*

5. Bursulaya, B.D., Totrov, M., Abagyan, R., Brooks, C.L. (2000) Comparative Study of Several Algorithms for Flexible Docking. *J. Comp. Chem.*(submitted).

Appendices

#> NA1	NA2	QMI	QMX	MI	MX	ID	SC	pP	DE
erbb2tk_human	1fgk_a20	3	260	26	303	37.8	116.34	23.05	"fgf receptor 1"
erbb2tk_human	1fmk_m15	5	256	191	434	39.3	105.66	20.62	"tyrosine-protein kinase src (equiv. To 2src"
erbb2tk_human	2ptk_m23	5	256	192	435	39.3	105.41	20.56	"tyrosine-protein kinase transforming protein src"
erbb2tk_human	1ir3_a19	2	254	21	284	37.8	104.92	20.45	"insulin receptor tkd"
erbb2tk_human	1irk_m21	2	254	21	284	38.1	106.52	20.81	"insulin receptor tkd mutant (c981s, y984f)"
erbb2tk_human	3lck_m17	1	256	16	267	35.2	100.38	19.41	"proto-oncogene tyrosine-protein kinase"
erbb2tk_human	1ad5_a26	3	265	187	435	32.6	98.22	18.92	"haematopoetic cell kinase hck"

Fig 1.
All the tyrosine kinase domains with known three-dimensional structure with substantial similarity (about 30%) to ErbB2 domain.

```
# Consensus                R...VLG      #G~V       ^I.#P~..         VA#K#L+~~.~.K...-
erbb2tk_human_   ----------RKVKVLGSGA--FGTVYK----GIWIPDGENVKIP--VAIKVLRENTSPKANKE
1fgk_a           ELPEDPRWELPRDRLVLG---KPLGQV--VLAEAIGLPNRV-----TKVAVKMLKSDATEKDLSD
#1fgk_a          _______________EEEE   EE___E  EEEEEE_____E      EEEEEE_______HHHHHH
##1fgk_a         _____X_____________   ______  _____________      _____B_____________

# Consensus      ##~E#~#M..#G      ~#..LLG#C.~      #~##.~.#..G~L.-~#...         .L^S.
erbb2tk_human_   ILDEAYVMAGVGSP---YVSRLLGICLTST---VQLVTQLMPYGCLLDHVRENRG-----RLGSQ
1fgk_a           LISEMEMMKMIG--KHKNIINLLGACTQ--DGPLYVIVEYASKGNLREYLQAR--RPPEEQLSSK
#1fgk_a          HHHHHHHHHHH_  _______EEEEE__  ____EEEE_______HHHHHH__  ________HH
##1fgk_a         ___________   _____________   ___________________________ _________

# Consensus      DL#~.^.Q#A+GM~YL.~.+.#HRDLAARNVLV.~.N.#KI.DFGLA       I~~.-Y
erbb2tk_human_   DLLNWCMQIAKGMSYLEDVRLVHRDLAARNVLVKSPNHVKITDFGLARLLD--IDETEYHA----
1fgk_a           DLVSCAYQVARGMEYLASKKCIHRDLAARNVLVTEDNVMKIADFGLA----RDIHHIDY--YKKT
#1fgk_a          HHHHHHHHHHHHHHHHHH_____________EEE_____EEE_______    _________  ____
##1fgk_a         _________________________A_____________________    _________  p___

# Consensus      ~~G+#P#KWMA#E^##.R.%THQSDVWS%GV.#WE##T#G^.PY.G#P#.E##.LL..G~R#..P
erbb2tk_human_   DGGKVPIKWMALESILRRRFTHQSDVWSYGVTVWELMTFGAKPYDGIPAREIPDLLEKGERLPQP
1fgk_a           TNGRLPVKWMAPEALFDRIYTHQSDVWSFGVLLWEIFTLGGSPYPGVPVEELFKLLKEGHRMDKP
#1fgk_a          ___________HHHHHH____HHHHHHHHHHHHHHHHH___________HHHHHHHHH_________
##1fgk_a         ____________________________________________________________________

# Consensus      ..CT.-#YM#M..CW.#..~~RP.F+~LV~-#~R##...~
erbb2tk_human_   PICTIDVYMIMVKCWMIDSECRPRFRELVSEFSRMARDPQR-----
1fgk_a           SNCTNELYMMMRDCWHAVPSQRPTFKQLVEDLDRIVALTS------
#1fgk_a          ____HHHHHHHHHH__________HHHHHHHHHHHHHHH__
##1fgk_a         ____________________________________________

# erbb2tk_human_ 1fgk_a nID 98 Lmin 265 ID 37.0 % Score 104.27 Sim 48.80 Gap 25.05
nOverlap 242 pP 20.54
#MATGAP gonnet 2.4 0.15
```

Figure 2. The alignment used to build the model by homology. The alignment was generated by the sequence - structure alignment algorithm which takes the secondary structure and accessibility into account. The gaps have been expanded to allow for modeling of short loops.

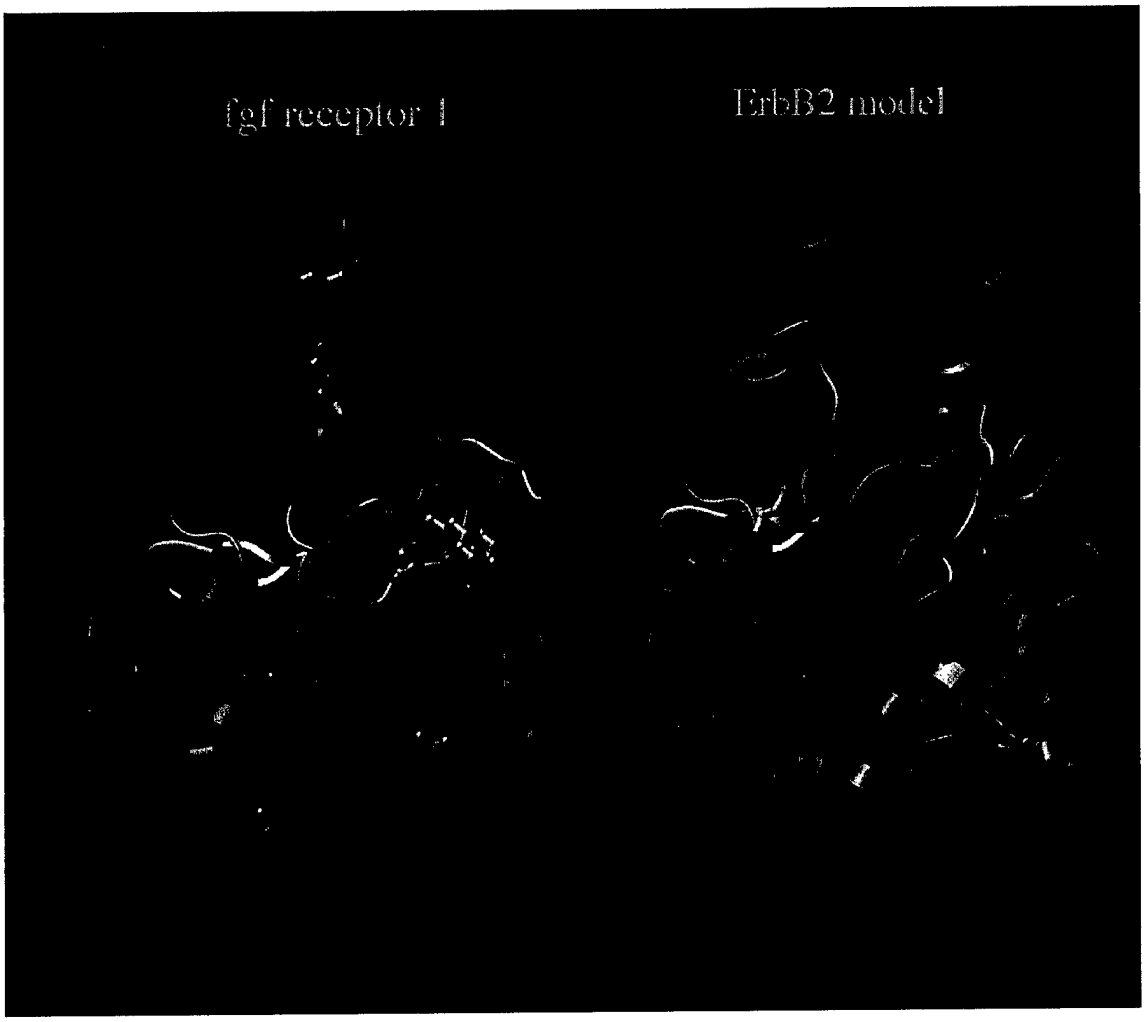
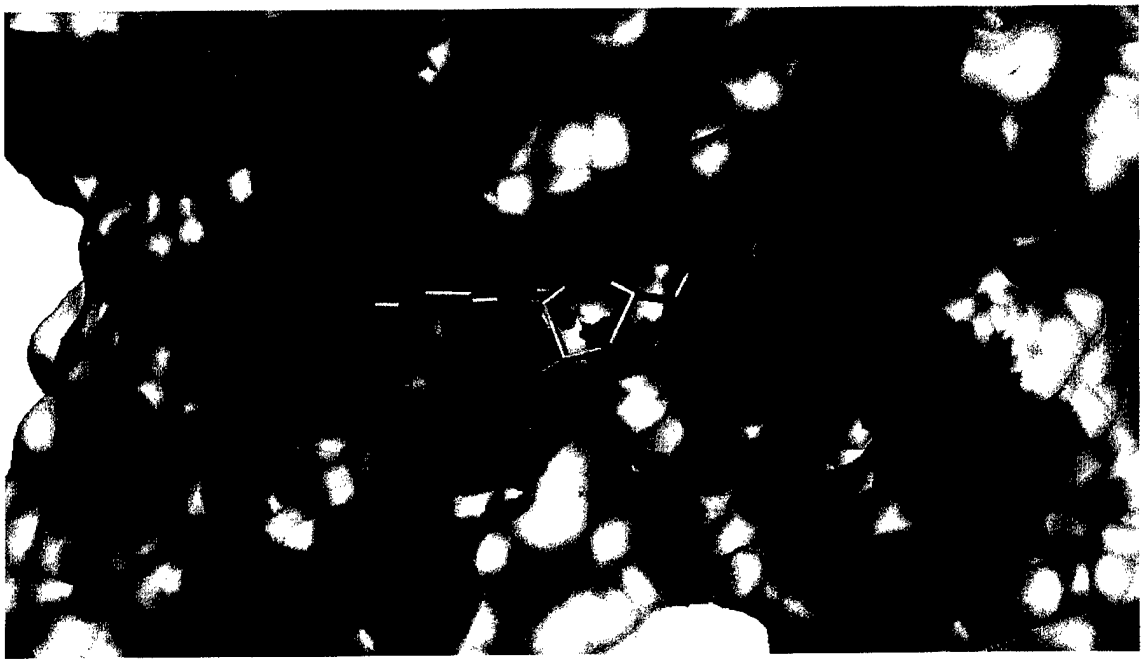


Figure 3. A side-by-side view of the template fibroblast growth factor receptor tyrosine kinase domain and the ErbB2 model. In yellow are shown the regions of sequence identity.

Fig. 4. The binding site of the ErbB2 model built using the SRC kinase domain template. The molecular surface of the site is colored by physico-chemical properities: hydrophobicity (green), hydrogen bonding donor (blue), hydrogen bonding acceptor (red).

Fig. 5 The binding site of the template tyrosine kinase domain (SRC kinase). This site is compared with the model of ErbB2 shown on Figure 4. The color code is the same as for Figure 4.



Rational discovery of novel nuclear hormone receptor antagonists

Matthieu Schapira*†, Bruce M. Raaka‡, Herbert H. Samuels‡, and Ruben Abagyan*†§

*Structural Biology, Skirball Institute of Biomolecular Medicine, and ‡Division of Molecular Endocrinology, Departments of Medicine and Pharmacology, New York University School of Medicine, 550 First Avenue, New York, NY 10016

Edited by Peter G. Schultz, The Scripps Research Institute, La Jolla, CA, and approved December 7, 1999 (received for review October 20, 1999)

Nuclear hormone receptors (NRs) are potential targets for therapeutic approaches to many clinical conditions, including cancer, diabetes, and neurological diseases. The crystal structure of the ligand binding domain of agonist-bound NRs enables the design of compounds with agonist activity. However, with the exception of the human estrogen receptor-α, the lack of antagonist-bound "inactive" receptor structures hinders the rational design of receptor antagonists. In this study, we present a strategy for designing such antagonists. We constructed a model of the inactive conformation of human retinoic acid receptor-α by using information derived from antagonist-bound estrogen receptor-α and applied a computer-based virtual screening algorithm to identify retinoic acid receptor antagonists. Thus, the currently available crystal structures of NRs may be used for the rational design of antagonists, which could lead to the development of novel drugs for a variety of diseases.

Members of the nuclear hormone receptor (NR) family are under the control of a wide variety of hormones and ligands, such as steroids, retinoids, thyroid hormone, 1,25-dihydroxy-vitamin D3, and prostanoids. Many of these NRs are potential targets for the therapy of a variety of diseases: antagonists of estrogen receptor-α (ERα) (e.g., tamoxifen) are clinically used for the treatment of breast cancer (1) whereas retinoic acid receptor (RAR) agonists and antagonists block the growth of a number of neoplastic cells including breast tumor cells (2, 3). Agonists for retinoid X receptors (RXRs) and peroxisome proliferator-activated receptor γ (PPARγ) are potential candidates for use in the treatment of cancer and diabetes (PPARγ is the receptor for the antidiabetic drug thiazolidinedione) (4–7), whereas Nurr1 ligands may be useful for treatment of Parkinson's disease (8). Thus, designing molecules that selectively activate or inhibit specific NRs is of considerable biological significance and will likely have the potential for use in important clinical applications.

The crystal structures of the ligand binding domain (LBD) of many members of the NR family recently have been solved, and the ligand-dependent structural changes involved in transcriptional activation have been clarified, enabling the structure-based design of specific agonists (9, 10). Recent studies on ERα also have shed light on the LBD structural changes mediated by NR antagonists (11, 12): ERα agonists and antagonists superimpose well and engage in a very similar network of hydrophobic and electrostatic contacts with the receptor. However, in the agonist-bound conformation, the C-terminal helix H12 sits like a lid on top of the ligand (11) (a similar observation was made for virtually all of the NR LBD structures solved so far; ref. 9). In contrast, the two ERα antagonists present a protruding arm that is not compatible with the "closed lid" conformation (11, 12) (Fig. 1*A*). As a result, helix H12 is pushed away from the ligand binding site and relocates in the coactivator-binding pocket of the receptor (Fig. 1*B*) (11). Moreover, the LxxML motif (where L is a leucine, M a methionine, and x any residue) of the ERα helix H12 mimics, and probably competes with, a LxxLL helical peptide found in a wide variety of coactivator proteins. The alignment of the LBD of various NRs (13) suggests that a common structural mechanism would be for the antagonists to induce the relocation of helix H12 into the hydrophobic coactivator-binding groove of the receptor. The observation that the progesterone receptor antagonist RU486 superimposes with the natural hormone progesterone, but presents a protruding arm similar to that of tamoxifen (14, 15) provides support for the universality of this mechanism of antagonistic activity.

Our goal in this study is to provide further evidence for this hypothesis by building a model of the antagonist-bound conformation of RARα, a NR that plays an important role in the differentiation and proliferation of a wide variety of cell types and for which only the agonist bound conformation is known (16–18), and to rationally and rapidly identify new antagonists for this receptor. We built a model of the antagonist-bound structure of RAR, based on the ERα/tamoxifen complex (12). The model was used for the virtual screening of a database of $\approx$150,000 available compounds, and antagonist candidates were tested *in vitro*. Two novel antagonists and a novel agonist were discovered. The ligands were specific for RAR, confirming the validity of our model and the potential therapeutic application of our strategy.

Materials and Methods

Building of the Model of Antagonist-Bound RAR. A helical peptide PLIREMLENP corresponding to helix H12 of RARγ was docked into the putative coactivator binding pocket of another RARγ molecule. We hypothesized that the IxxML motif contacts the coactivator binding site of the receptor, and an automatic docking procedure was carried out toward this site, with flexible protein and peptide side chains, according to a biased probability Monte Carlo energy minimization procedure (19, 20). Two critical features of the interaction between the LBDs of NRs and their coactivators were used to carry out the docking: (*i*) The "charge clamp," initially observed in the complex between SRC-1 and peroxisome proliferator-activated receptor γ (21), where a conserved glutamate (E414 in RARγ) and lysine (K246 in RARγ) at opposite ends of the hydrophobic cavity of the receptor contact the backbone of the coactivator's LxxLL box, enabled the orientation of the helical peptide. (*ii*) The finding that the leucines of the LxxLL motif of SRC-1 are buried in the hydrophobic cavity of the receptor determines which side of the helix faces the receptor. Here, the isoleucine, methionine, and leucine of the IxxML motif were buried in the binding site of RARγ. Loose distance restraints were set between the charge

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: NR, nuclear hormone receptor; RAR, retinoic acid receptor; ER, estrogen receptor; LBD, ligand binding domain; RXR, retinoid X receptor; CAT, chloramphenicol acetyltransferase.

†To whom reprint requests should be addressed. E-mail: abagyan@scripps.edu or schapira@saturn.med.nyu.edu.

§Present address: Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, MB-37, La Jolla, CA 92037.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

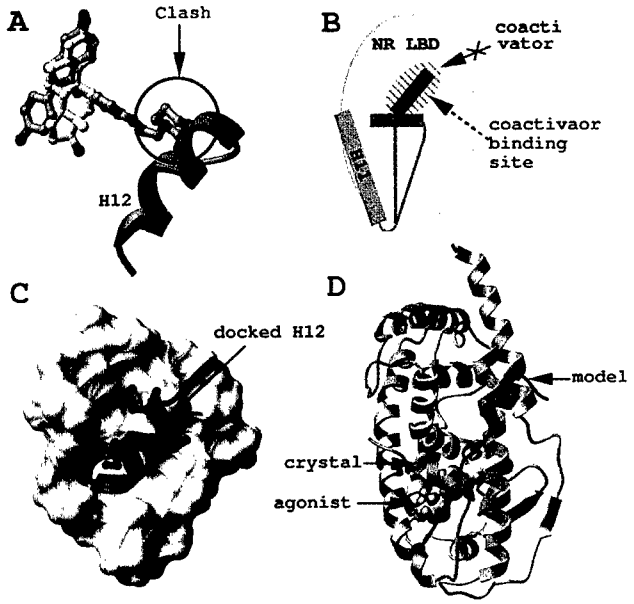


Fig. 1. Modeling of the antagonist-bound structure of RAR. Agonist (white) and antagonist (cyan) superimpose in the binding pocket of ERα, but the antagonist presents an additional protruding arm that pushes helix 12 (H12, green) away (*A*). As a result, H12 relocates in the coactivator binding pocket of the receptor (H12, red) (*B*). Based on the ERα structure, helix H12 of RARγ (red) was docked to the coactivator binding pocket of the RARγ-LBD (critical hydrophobic residues are displayed in magenta) (*C*), and the C terminus of the protein was remodeled from its agonist-bound conformation (green) to its antagonist-bound conformation (red) (*D*).

clamp of the receptor (21) (i.e., E414 and K246) and backbone nitrogen and oxygens of the peptide (nitrogen of the isoleucine on one end, and carbonyl of the methionine, leucine, and asparagine in the MLEN motifs, respectively). The energy of the complex was minimized in the internal coordinate space by using the modified ECEPP/3 potentials. The subset of the variables minimized with the ICM method (19, 20, 22, 23) included the side chains of the receptor, six positional variables of the helix, and the side-chain torsion angles of the helix.

After the ICM docking procedure, we built a model of antagonist-bound RARγ. The structure of the receptor was kept rigid but for the side chains and backbone of the 25 C-terminal residues (corresponding to the last 10 residues of helix H11, the loop from H11 to H12, and H12), and for the side chains of the putative coactivator binding site (within 6 Å of the previously docked helical peptide). Tethers then were set between the C terminus of the receptor and the corresponding residues of the docked helical peptide, and the energy of the receptor was minimized by a stochastic global energy optimization in the internal coordinate space (22, 23).

The last step was, from the resulting model of antagonist-bound RARγ, to derive the structure of the antagonist-binding pocket of RARα: the three nonidentical residues in the vicinity of the binding pocket (A234, M272, and A397) were changed to the RARα isoform (S234, I272, and V397, respectively) and energy-minimized. Another possibility would have been to introduce the mutations before remodeling the C terminus of the receptor. We preferred to proceed as described here to preserve the integrity of the receptor during the critical remodeling of the C-terminal end.

Receptor-Ligand Docking. An initial docking was carried out with a grid potential representation of the receptor and flexible ligand (24). The resulting conformation then was optimized with a full atom representation of the receptor, flexible receptor side chains, and flexible ligand, by an ICM stochastic global optimization algorithm as implemented in the MolSoft ICM 2.7 program (23, 24).

Screening of a Virtual Library of Compounds. The flexible-ligand/grid-potential-receptor docking algorithm (23, 24) was carried out automatically on the Available Chemicals Directory library of 153,000 available chemical compounds (MDL Information Systems, San Leandro, CA). The screening took less than a month on 10 194-MHz IP25 processors. Each compound was assigned a score, according to its fit with the receptor, which took into account continuum as well as discreet electrostatics, hydrophobicity, and entropy parameters (25). The distribution of the compounds according to their score is presented at http://abagyan.scripps.edu/PNAS/MS2000/. All compounds scoring better (i.e., lower) than −32 were screened further for the number of hydrogen bonds engaged with the receptor. The 134 compounds that made at least two hydrogen bonds with the receptor were preselected. The 609 compounds scoring better than −37 also were preselected, regardless of the hydrogen bonding network. This preselection pool then was further minimized with a full atom representation of the receptor, as described above. The quality of the fit of the 500 best-scoring compounds then was visually estimated, and 32 compounds were selected for biological testing. These compounds are not necessarily the ones with the best final scores, but the ones we thought, after careful visual inspection, presented the best characteristics, such as Van der Waals fit or hydrogen bonding (see http://abagyan.scripps.edu/PNAS/MS2000/).

It occurred to us that during the selection by the MolSoft virtual screening procedure, it was preferable to set up an initial cut-off value poorly selective (i.e., −32) to recover a large pool of preselected compounds and to apply to this pool subsequent screens specific for the system, such as number of hydrogen bonds (used here) or presence of a hydrogen bond acceptor (for example) at a specific point of space. As a result, we derived the value −32 as a good initial threshold (this value generates an initial pool of 3,000–4,000 compounds).

Biological Activity of the Antagonist and Agonist Candidates. HeLa cells were transfected by calcium phosphate precipitation using 1 μg of the Gal4-responsive chloramphenicol acetyltransferase (CAT) reporter pMC110 and 1 μg of Gal4-hRARα-LBD or 1 μg of Gal4-hRXRβ-LBD. Studies also were performed with the three wild-type hRAR isoforms (hRARα, hRARβ, and hRARγ) by using a ΔMTV-IR-CAT reporter as described (26, 27). Cell cultures were supplemented with indicated ligands immediately after addition of the calcium phosphate/DNA precipitate. Media and ligands were replaced after 24 h, and cells were harvested and essayed for CAT activity 24 h later.

Results

Modeling of the RAR Antagonist Binding Pocket. The x-ray structure of RARγ bound to the agonist all-trans RA is available (18); however, the conformation of the receptor bound to an antagonist is not known. We used the observations made from the structure of ERα bound to an agonist, 17β-estradiol (11), and two antagonists, tamoxifen and raloxifene (11, 12), to build a model of antagonist-bound RAR (Fig. 1 *A* and *B*). We docked helix H12 of RARγ into the putative coactivator binding pocket of the receptor as described (27) (see *Materials and Methods* for details) (Fig. 1*C*) and remodeled the 25 C-terminal residues, starting near the end of helix 11, through an extensive global energy minimization procedure (Fig. 1*D*).

Docking of Known RAR Antagonists into the Modeled Receptor. A few RAR antagonists have been described in the literature; and

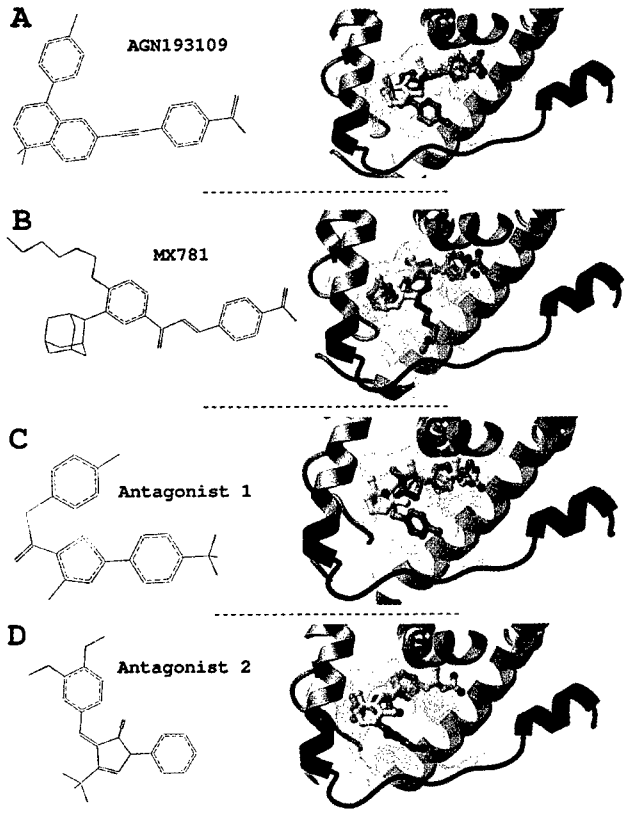


Fig. 2. RAR antagonists. Two known antagonists (*A* and *B*) and two novel antagonists (*C* and *D*). (*Left*) Chemical structure. (*Right*) Conformation docked into the receptor (part of the receptor is displayed as a ribbon representation, and the binding pocket boundary is displayed in yellow). Cyan, carbons; red, oxygen; blue, nitrogen; magenta, fluorine; yellow, sulfur. Hydrogens are not represented for clarity.

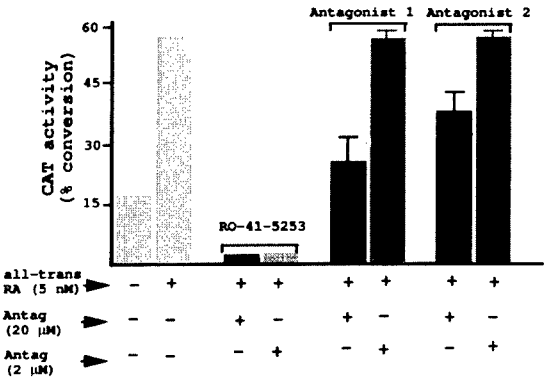


Fig. 3. Functional assays of the novel antagonists. HeLa cells were transfected with a Gal4-hRARα-LBD expression vector and a Gal4-CAT reporter gene (results were similar in studies using the three hRAR isoforms). The cells were incubated with 5 nM all-trans RA to stimulate CAT activity, and the effect of each antagonist on inhibiting CAT was examined at 2 and 20 μM concentration (the known antagonist RO-41–5253 was used as a positive control).

several of them are serious candidates for cancer therapy (2, 28). A well-characterized ligand is AGN193109, which inhibits the three RAR isoforms at nanomolar concentrations (29). Another very potent antagonist is MX781, which is effective against ERα-positive and -negative breast cancer cells, with no apparent toxicity (2). The activity of these two ligands has been presented in detail, but no structural information has been reported on their mode of interaction with the receptor. We built a model of RARγ complexed either with AGN193109 or MX781, by using our flexible docking algorithm (24) (Fig. 2 *A* and *B*). In both cases, the antagonist superimposed with the agonist all-trans RA. As observed for ERα, the antagonists also presented a protruding arm, which was absent in RAR agonists. Very importantly, this protruding arm coincided exactly with the single opening in the ligand binding pocket of our modeled receptor, generated by the displacement of helix H12 (Fig. 2 *A* and *B*), and made stabilizing hydrophobic contacts with the protein. It is very unlikely that this perfect fit, observed for both antagonists, was fortuitous. On the contrary, this feature mimics the inactivation mechanism revealed by the crystal structure of ERα bound to tamoxifen and raloxifene. Therefore, our docking results of AGN193109 and MX781 very strongly suggest that: (*i*) the structural mechanisms of antagonist activity for ERα are shared by other NRs, and (*ii*) our model of the RAR antagonist binding pocket could be used to design novel antagonists.

Screening of a Virtual Library and Discovery of Novel RAR Antagonists. High throughput functional screening currently is the most used method for the discovery of receptor-specific ligands. Although efficient, it requires the physical availability and management of hundreds of thousands of chemical compounds. In the present work, we used a virtual library composed of the predicted structure of more than 150,000 available compounds (see *Materials and Methods*). Each compound was automatically docked in a grid representation of the modeled RARα antagonist binding pocket. Five grid potentials carried information on the shape, hydrophobicity, electrostatics, and hydrogen-bonding availability of the receptor, and enabled a rapid docking simulation (24, 25). RARα was selected over the other two isoforms (RARβ and RARγ) because recent data suggests it could be a medically more relevant target (28). After an automatic selection procedure with flexible ligands, and optimization of the selected candidates with flexible protein side chains (see *Materials and Methods* for details), 32 compounds were considered as potential antagonists of RARα and ordered.

To test these compounds *in vitro*, HeLa cells were transfected with a Gal4-hRARα-LBD expression vector and a Gal4-CAT reporter gene (26). Studies also were performed with the three wild-type hRAR isoforms and a ΔMTV-IR-CAT reporter (26, 27). These gave similar results as those found with Gal4-hRARα-LBD (data not shown). The cells were incubated with all-trans RA to stimulate CAT activity, and the effect of each antagonist candidate on inhibiting CAT stimulation by all-trans RA was examined. Possible toxicity of the compounds was deduced from the amount of cellular protein extract after 2 days of incubation. Two antagonist candidates inhibited CAT activity by 55% and 33% at 20 μM with no apparent toxicity (Fig. 3). The Gal4-hRARα activity illustrated in Fig. 3 was equivalent for the other two RAR isoforms (data not shown). No inhibition was observed when CAT expression was under the control of a Gal4-mRXRβ-LBD fusion construct, indicating that: (*i*) the antagonists are specific for RAR, and (*ii*) the inhibition is caused by an interaction with the Gal4-RAR-LBD fusion protein and does not result from some nonspecific effect on CAT activity (data not shown).

The two RAR antagonists dock into the ligand binding pocket of the receptor (Figs. 2 *C* and *D* and 4). As observed for AGN193109 and MX781, they fit in the same binding pocket as the natural agonist all-trans RA, but present an additional arm, which protrudes out of the pocket. Antagonist 1 has a tri-fluoro group where the retinoid receptor ligands usually carry a carboxylate group (in antagonist 2, the corresponding domain is truncated). In our model, antagonist 2 engages in a hydrogen bond with Ser-234 of the hRARα (Fig. 4*B*). However, the S234A

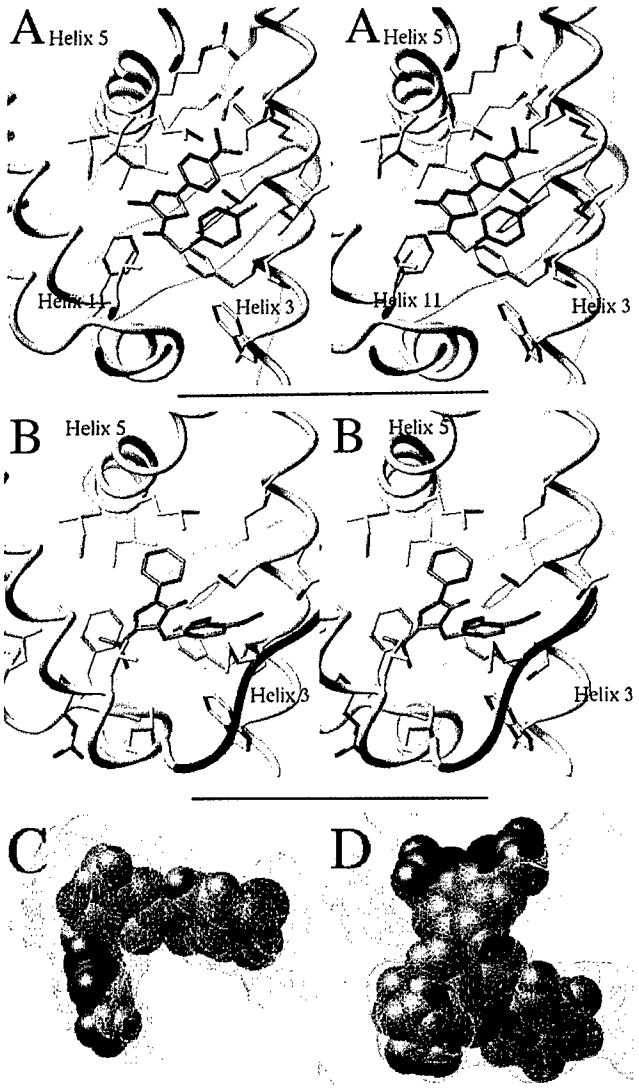


Fig. 4. Novel RAR antagonists. (*A* and *B*) Stereo representation of antagonists 1 and 2 docked into the binding site of the receptor. The ligands make extensive hydrophobic interactions with residues from helix 3, helix 5, and helix 11. Antagonist 2 (*B*) is engaged in an additional hydrogen bond with Ser-234 of helix 3 and contacts the remodeled C terminus (red) at Pro-405. (*C* and *D*) The fit of antagonists 1 and 2 into the receptor binding pocket is shown.

mutation in the other two isoforms does not alter the ligand antagonist activity, suggesting that this hydrogen bond is not essential for the interaction. An obvious way to increase the affinity of these antagonists would be to substitute the tri-fluoro group by a carboxylate in antagonist 1 or elongate and add a carboxylate to antagonist 2, which would result in more stabilizing interactions with two conserved arginines of the receptor. However, the purpose of this work is to provide evidence that the rational design of antagonists from the model of the inactive receptor is feasible and not to optimize the affinity of the compounds. The *in vitro* functional assays provide evidence that our modeling scheme is relevant and can be used to design novel antagonists of NRs.

We applied the same strategy to discover agonists, by using the crystal structure of the active conformation of RARγ (18), and could discover three novel agonists 10–25% active at 200 nM and fully active at 20 μM, of 30 compounds tested (data not shown).

Screening of a Database of Known Ligands. To assess the quality of our setup of the ICM screening algorithm (23), we built a small virtual database made up of antagonists and agonists for different members of the NR family (Table 1). We screened this database with our model of antagonist-bound RAR, as we did for the Available Chemicals Directory database. The screening was repeated four times, to test the reproducibility of our method. Table 1 shows that for each ligand the score varies a lot from one screening to the other. This finding reflects the generation of different ligand conformations from one docking simulation to another (data not shown) and represents the limitation of our method, as discussed below.

Table 1 lists as "selected" the ligands that met with the criteria for preselection and final inspection during the Available Chemicals Directory screening (i.e., score better than −37 or score better than −32 and at least two hydrogen bonds with the receptor; see *Materials and Methods* for details). Seven of the nine known RAR ligands (i.e., ≈80%) and one of the six non-RAR ligands (i.e., ≈16%) were selected. The fact that RAR agonists, as well as antagonists, produced good scores was expected, because the binding pocket used for the screening is equivalent to the agonist binding pocket, with an additional opening generated by the remodeling of the C terminus of the receptor. The two false negatives, AGN193836 and Ro415253, were missed because of steric clashes, as discussed below. Antagonist 1 was not found either, reflecting its rather low affinity for the receptor. It is important to underline here that we do not expect to detect all of the true binders. The algorithm was rather designed to minimize the number of false positives, which correlates with the number of unnecessary *in vitro* experiments (25).

In that respect, the presence of one false positive of six nonbinders could be alarming, because such a ratio would represent about 25,000 false positives of a database of 150,000 compounds. However, the binding pockets of the NRs represented in this database are close in size and shape; as a result, the database used for this benchmark was composed of molecules presenting strong similarities with RAR ligands. Therefore, we believe this ratio is not representative. The fact that we needed to test only 32 molecules to discover three novel RAR ligands confirms this assumption.

Next, we tried to address why some ligands, such as Ro415253, were repeatedly missed by our screening algorithm (Ro415253 was still not selected after 10 docking simulations, data not shown). We hypothesized that the ligand could not fit into the potential maps generated from our model and carried out a docking simulation with a full atom representation of the receptor, according to a Monte Carlo energy minimization of the complex, with both flexible ligand and flexible receptor side chains (24). This docking simulation produced a solution were the ligand fits into the binding pocket; the core of the ligand (from the carboxylate to the internal sulfone) superimposes with agonists such as all-trans RA, whereas the alkyl arm sticks out of the pocket, as previously described for the other antagonists (data not shown). The conformation of several receptor side chains was modified during the docking simulation, to accommodate the size of the ligand, and this solution would not have been found with rigid side chains. This finding suggests that Ro415253 could not fit into the potential maps generated from the original receptor conformation, which we used for the screening. We generated a new series of potential maps from the optimized receptor structure and screened the small database of known ligands with these maps four times as above (Table 1). The score assigned to Ro415253 was twice lower (i.e. better) than the threshold. Surprisingly, this new series of potential maps totally eliminated the presence of both false positive and false negative (all RAR ligands and only RAR ligands were selected).

Table 1. Control screening of known NR ligands

Ligand	Activity	Score 1	Score 2	Score 3	Score 4	Selected	Binding	References
First series								
AGN193836	RAR_agonist	−19.9	−9.04	−20.6	−19.7	−	+	(33)
ATRA	RAR pan-agonist	−46.4	−41	−41.7	−41.	+	+	(34)
Ro415253	RAR_antagonist	−25.5	−22.	−28.3	−28.6	−	+	(28)
MX781	RAR antagonist	−28.	−23.9	−27.1	−36.4	+	+	(2)
CD2366	RAR pan-antagonist	−28.5	−23.3	−30.9	−32.3	+	+	(34)
Targretin	RXR pan-agonist	−17.9	−18.1	−19.1	−18.6	−	−	(4)
SR11203	RXR pan-agonist	−27.5	−27.	−27.	−27.2	−	−	(34)
Tamoxifen	ER modulator	−29.3	−27.5	−29.8	−28.3	−	−	(23)
Raloxifene	ER modulator	−23.4	−20.8	−26.7	−34.6	+	−	(22)
RU486	Progest Rec antag.	−21.2	−21.3	−21.4	−21.3	−	−	(25)
9cisRA	RAR/RXR agonist	−32.5	−32.6	−32.9	−16.9	+	+	(34)
AGN193109	RAR pan-antagonist	−39.2	−56.	−57.4	−39.4	+	+	(29)
AGNpartia	RAR partial agonist	−54.4	−54.3	−49.5	−29.1	+	+	(29)
Am580	RAR_agonist	−34.2	−34.4	−34.8	−34.5	+	+	(34)
EM652	ER antagonist	−27.	−27.4	−21.7	−28.8	−	−	(35)
Antagonist 1	Novel RAR antag.	−28.5	−28.1	−28.7	−28.8	−	+	(35)
Antagonist 2	Novel RAR antag.	−27.6	−38.9	−40.2	−26.3	+	+	(35)
Second series								
AGN193836	RAR_agonist	−37.2	−36.5	−36.7	−35.3	+	+	(33)
ATRA	RAR pan-agonist	−51.7	−52.6	−51.8	−52.0	+	+	(34)
Ro415253	RAR_antagonist	−28.9	−24.4	−39.0	−46.6	+	+	(28)
MX781	RAR antagonist	−45.3	−48.0	−40.2	−45.6	+	+	(2)
CD2366	RAR pan-antagonist	−50.7	−50.8	−29.3	−29.3	+	+	(34)
Targretin	RXR pan-agonist	−25.4	−23.0	−22.2	−31.0	−	−	(4)
SR11203	RXR pan-agonist	−28.2	−22.7	−22.1	−27.5	−	−	(34)
Tamoxifen	ER modulator	−26.4	−24.6	−30.3	−23.4	−	−	(23)
Raloxifene	ER modulator	−15.6	−23.7	−18.4	−17.4	−	−	(22)
RU486	Progest Rec antag.	−21.4	−20.6	−20.3	−20.1	−	−	(25)
9cisRA	RAR/RXR agonist	−38.8	−39.5	−33.5	−38.7	+	+	(34)
AGN193109	RAR pan-antagonist	−55.1	−55.5	−41.2	−54.8	+	+	(29)
AGNpartia	RAR partial agonist	−61.4	−61.3	−61.4	−61.0	+	+	(29)
Am580	RAR_agonist	−46.6	−47.2	−46.6	−46.5	+	+	(34)
EM652	ER antagonist	−26.3	−23.1	−23.7	−27.3	−	−	(35)
Antagonist 1	Novel RAR antag.	−32.1	−32.1	−31.7	−31.6	+	+	(35)
Antagonist 2	Novel RAR antag.	−33.3	−29.7	−33.8	−33.8	+	+	(35)

First series: A similar screening as the one performed on the ACD database was carried out four times on a small database made of known RAR antagonists, agonists, as well as ligands for other NRs and the two novel RAR antagonists. The ligands that met at least once with the criteria for selection used during the ACD screening are listed as Selected. The ligands that are experimentally binding to RAR are listed as Binding. Second series: Screening of known ligands after adjustment of the receptor's binding pocket conformation. The RAR antagonist Ro415253 was docked into our model of antagonist-bound RAR with flexible receptor side chains and ligand. The resulting receptor conformation was used for a novel screening.

Discussion

In this study, we presented a strategy for the discovery of antagonists, as well as agonists, for NRs, which are very important targets for drug design. An important aspect of our approach was to exclude any preconceived pharmacophore bias from our database screening. Most drug design strategies impose chemical constraints on the selected molecule to conserve the functional groups believed to be most important in existing ligands, preventing the discovery of novel ligand types. In the present work, we avoided pharmacophore constraints thanks to a robust flexible docking program and scoring function: the only filters used for screening were a good fit with the receptor and reasonable bioavailability parameters (30). As a result, we discovered novel original ligands that could be further optimized into potent RAR-selective antagonists and agonists.

A limitation of our method, which leaves room for further improvement, is that a compromise must be made between the time allocated for each ligand (less than 2 min on one processor here) and the reliability of the sampling of the conformational space. Indeed, Table 1 shows that four runs for each ligand are necessary to minimize efficiently missed hits (the remaining missed positives were not selected because of inappropriate receptor side-chain conformations and not because of an insufficient sampling). Improvement of the computing power, the docking algorithm, and the scoring function all could result in a more robust virtual database screening.

Another drawback is that the conformation of the receptor is not necessarily unique, but can vary from one ligand to another. As a result, a ligand that fits in receptor conformation A will never be found if receptor conformation B is used for the screening. The case of Ro415253 illustrates this issue well: this known antagonist was never selected, even after 10 trials, because the binding pocket used for the screening was too narrow. The potential maps used for the screening have a smoother van der Waals profile than the atomic representation of the receptor; as a result, the maps are more tolerant regarding steric clashes with the ligand. However, the degree of tolerance is limited and cannot accommodate important conformational changes of the receptor side chains (or backbone, obviously). When new potential maps generated from a model of RAR bound to Ro415253 were used for screening, the three RAR ligands missing from the first screening were selected (Table 1). This finding confirms that the initial conformation of the

receptor prevented the selection of, or reduced the chances of selecting, some known RAR ligands. The false positive raloxifene (Table 1) was making extensive van der Waals interactions with the narrow RAR binding pocket, which compensated for the lack of stabilizing electrostatic interactions. However, in the new conformation of the receptor (Table 1), the binding pocket is wider and the fit not as tight. As a result, raloxifene was not selected. This observation emphasizes, if necessary, that virtual screening is very sensitive to the conformation of the receptor.

In that respect, it is interesting to note that the topology of the remodeled C-terminal loop is probably not unique, and that the conformation used to generate the receptor potential maps was one among many others. It is therefore legitimate to wonder whether novel antagonists could not be discovered as efficiently from a structure of the receptor where the C terminus, instead of being remodeled, was truncated. This brings up a fundamental question: is the role of antagonists only to antagonize the "closed lid" conformation where helix H12 sits on top of the ligand binding pocket, or are they also stabilizing the inactive conformation of the receptor? It is important to keep it mind that the C-terminal tail of RAR (as well as for other NRs) is a very dynamic entity when no ligand is bound to the receptor and probably oscillates between active and inactive conformations. Once bound in the ligand binding pocket, agonists contact the H12 helix and lock the receptor in its coactivator-binding conformation. Likewise, it is reasonable to speculate that antagonists would contact the C-terminal tail of the receptor and stabilize the inactive state. However, it is probable that the conformation of the receptor varies from one ligand to another; indeed, recent results on ERα show that different ligands induce distinct conformational change of the receptor (31). We used the crystal structure of ERα bound to tamoxifen to build our model of inactive RAR and could find two specific antagonists, one of which contacts the remodeled tail of the receptor. Although the conformation we used for the C-terminal tail was probably not the only possible one, we believe that its presence was important to bias the screening toward compounds that actually do contact the flexible arm of RAR, as well as to impose a reasonable boundary on the antagonist binding pocket, and prevent the ligands from drifting out of the pocket during the docking simulations.

An important point was to demonstrate that we could discover novel antagonists for a NR other than ERα, provided that the structure of the agonist-bound active form of the protein was known. Rational design of ligands from a model of a receptor is thought by many to yield very low success rates. The present study demonstrates that this strategy can be successfully undertaken with appropriate biological systems and robust modeling tools. Moreover, targeting models of diverse members of the NR family could be further justified by the wealth of structural and sequence information (9, 13), as well as the finding that NR family members share similar mechanisms of transcriptional activation and inhibition (9).

The recent publication of the crystal structures of medically relevant receptor targets, such as peroxisome proliferator-activated receptor γ (21), RAR (18), RXR (32), ERα (11), or progesterone receptor (15), has created an exciting opportunity for the discovery of novel ligands. This study demonstrates that the rational design of both antagonists and agonists, by using computer-generated models based on these structures, is possible.

We thank M. Totrov for helpful discussion. We thank MolSoft LLC for making the latest version of the ICM program available for this research project. This research was supported by Department of Defense Grant DAMD179818133, a Kaplan Comprehensive Cancer Center grant from New York University Medical Center, National Institutes of Health Grant GM5541801, and Department of Energy Grant DEFG0296ER62268 to M.S. and R.A., and National Institutes of Health Grant DK16636 and New York State Empire Award C015710 to H.H.S.

1. Dees, E. C. & Kennedy, M. J. (1998) *Curr. Opin. Oncol.* **10,** 517–522.
2. Fanjul, A. N., Piedrafita, F. J., Al-Shamma, H. & Pfahl, M. (1998) *Cancer Res.* **58,** 4607–4610.
3. Shiohara, M., Dawson, M. I., Hobbs, P. D., Sawai, N., Higuchi, T., Koike, K., Komiyama, A. & Koeffler, H. P. (1999) *Blood* **93,** 2057–2066.
4. Bischoff, E. D., Moon, T. E., Heyman, R. A. & Lamph, W. W. (1998) *Cancer Res.* **58,** 479–484.
5. Mukherjee, R., Davies, P. J., Paterniti, J. R., Jr. & Heyman, R. A. (1997) *Nature (London)* **386,** 407–410.
6. Spiegelman, B. M. (1998) *Diabetes* **47,** 507–514.
7. Elstner, E., Muller, C., Koshizuka, K., Williamson, E. A., Park, D., Asou, H., Shintaku, P., Said, J. W., Heber, D. & Koeffler, H. P. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 8806–8811.
8. Zetterstrom, R. H., Solomin, L., Jansson, L., Hoffer, B. J., Olson, L. & Perlmann, T. (1997) *Science* **276,** 248–250.
9. Moras, D. & Gronemeyer, H. (1998) *Curr. Opin. Cell Biol.* **10,** 384–391.
10. Klaholz, B. P., Renaud, J. P., Mitschler, A., Zusi, C., Chambon, P., Gronemeyer, H. & Moras, D. (1998) *Nat. Struct. Biol.* **5,** 199–202.
11. Brzozowski, A. M., Pike, A. C., Dauter, Z., Hubbard, R. E., Bonn, T., Engstrom, O., Ohman, L., Greene, G. L., Gustafsson, J. A. & Carlquist, M. (1997) *Nature (London)* **389,** 753–758.
12. Shiau, A. K., Barstad, D., Loria, P. M., Cheng, L., Kushner, P. J., Agard, D. A. & Greene, G. L. (1998) *Cell* **95,** 927–937.
13. Wurtz, J. M., Bourguet, W., Renaud, J. P., Vivat, V., Chambon, P., Moras, D. & Gronemeyer, H. (1996) *Nat. Struct. Biol.* **3,** 87–94.
14. Cadepond, F., Ulmann, A. & Baulieu, E. E. (1997) *Annu. Rev. Med.* **48,** 129–156.
15. Williams, S. P. & Sigler, P. B. (1998) *Nature (London)* **393,** 392–396.
16. Fitzgerald, P., Teng, M., Chandraratna, R. A., Heyman, R. A. & Allegretto, E. A. (1997) *Cancer Res.* **57,** 2642–2650.
17. Giannini, G., Dawson, M. I., Zhang, X. & Thiele, C. J. (1997) *J. Biol. Chem.* **272,** 26693–26701.
18. Renaud, J. P., Rochel, N., Ruff, M., Vivat, V., Chambon, P., Gronemeyer, H. & Moras, D. (1995) *Nature (London)* **378,** 681–689.
19. Totrov, M. & Abagyan, R. (1994) *Nat. Struct. Biol.* **1,** 259–263.
20. Strynadka, N. C., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B. K., Kuntz, I. D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., *et al.* (1996) *Nat. Struct. Biol.* **3,** 233–239.
21. Nolte, R. T., Wisely, G. B., Westin, S., Cobb, J. E., Lambert, M. H., Kurokawa, R., Rosenfeld, M. G., Willson, T. M., Glass, C. K. & Milburn, M. V. (1998) *Nature (London)* **395,** 137–143.
22. Abagyan, R. & Totrov, M. (1994) *J. Mol. Biol.* **235,** 983–1002.
23. MolSoft (1998) ICM 2.7 *Program Manual* (MolSoft, San Diego).
24. Totrov, M. & Abagyan, R. (1997) *Proteins* Suppl. **1,** 215–220.
25. Totrov, M. & Abagyan, R. (1999) in *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, April 1999, Lyon, France (ACM Press, New York), pp. 37–38.
26. Qi, J. S., Desai-Yajnik, V., Greene, M. E., Raaka, B. M. & Samuels, H. H. (1995) *Mol. Cell. Biol.* **15,** 1817–1825.
27. Li, D., Desai-Yajnik, V., Lo, E., Schapira, M., Abagyan, R. & Samuels, H. H. (1999) *Mol. Cell. Biol.* **19,** 7191–7202.
28. Toma, S., Isnardi, L., Raffo, P., Riccardi, L., Dastoli, G., Apfel, C., LeMotte, P. & Bollag, W. (1998) *Int. J. Cancer* **78,** 86–94.
29. Chandraratna, R. A. (1998) *J. Am. Acad. Dermatol.* **39,** S149–S152.
30. Lipinski, C. A., Lombard, F., Dominy, B. W. & Feeney, P. J. (1997) *Adv. Drug Delivery Rev.* **23,** 3–25.
31. Paige, L. A., Christensen, D. J., Gron, H., Norris, J. D., Gottlin, E. B., Padilla, K. M., Chang, C. Y., Ballas, L. M., Hamilton, P. T., McDonnell, D. P. & Fowlkes, D. M. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 3999–4004.
32. Bourguet, W., Ruff, M., Chambon, P., Gronemeyer, H. & Moras, D. (1995) *Nature (London)* **375,** 377–382.
33. Teng, M., Duong, T. T., Klein, E. S., Pino, M. E. & Chandraratna, R. A. (1996) *J. Med. Chem.* **39,** 3035–3038.
34. Sun, S. Y., Yue, P., Dawson, M. I., Shroot, B., Michel, S., Lamph, W. W., Heyman, R. A., Teng, M., Chandraratna, R. A., Shudo, K., *et al.* (1997) *Cancer Res.* **57,** 4931–4939.
35. Tremblay, A., Tremblay, G. B., Labrie, C., Labrie, F. & Giguere, V. (1998) *Endocrinology* **139,** 111–118.

Protein-ligand docking as an energy optimization problem.

Maxim Totrov[1] and Ruben Abagyan[1,2]

[1]The Scripps Research Institute,

MB37 10550 North Torrey Pines Rd.,

San Diego, CA 92037, USA.

[2]Novartis Institute of Functional Genomics,

3115 Merryfield Row,

San Diego, CA 92121, USA.

Introduction

Formation of non-covalent complexes is an essential part of almost any biological process. Remarkable complexity of the biochemical machinery of the living organisms would have been impossible without the ability of the participating molecules to recognize each other among thousands of other compounds simultaneously present in any cell. Specific binding between molecules is crucial in catalysis, signal transduction, molecular transport mechanisms, and determines the pharmacological effect of many drugs.

Better knowledge of the nature of molecular recognition on the microscopic level is important for our understanding of the normal and pathological processes in the cell and may help in such practical applications as drug design. X-ray crystallography has revealed detailed atomic descriptions of many individual proteins, nucleic acids and small biological molecules, as well as a number of structures of complexes. The Protein Data Bank (PDB) (Bernstein et al. 1977), where solved protein 3D structures are deposited is growing by about 1000 new structures a year. Available structures of complexes can be analyzed to discover the basic interactions and principles of molecular recognition, while the individual structures can be used in the prediction of unknown or novel complexes. First attempts to predict molecular interactions and design novel ligand utilized hand-made physical models of receptor sites and ligands (Beddell *et al.* 1976). Since the manipulation of the systems containing hundreds or thousands of atoms is necessary to simulate the binding process, the progress in numerical computational approaches was essential for the advancement of the macromolecular association studies. Computer simulations of molecular recognition were first attempted more then twenty years ago (Kuntz *et al.* 1982). Considerable progress has been achieved in the recent years, but reliability and precision of the existing complex prediction methods is still far from ideal.

Molecular docking simulations

Prediction of the structure of a complex starting from the structures of individual molecules is commonly called molecular docking problem. Structures of the protein-ligand and especially protein-protein complexes often show remarkable shape complementarity on the interface, suggesting the idea that the docking algorithms should search for such matching surfaces. Early approaches such as the original DOCK (Kuntz *et al.* 1982) used exclusively this geometric criterion. Both components of the complex were assumed rigid and the docking procedure searched for favorable mutual orientation using "sphere matching" (DesJarlais *et al.* 1986), least-squares fitting of the surface patterns (Bacon and Moult 1992, Leach and Kuntz 1992), Fourier-transform (Katchalski-Katzir *et al.* 1992), distance-matrix matching (Helmer-Citterich and Tramontano 1994) or "geometric hashing" (Fischer *et al.* 1995). Purely geometric approaches demonstrated certain success in recombining the structures of protein-protein complexes when the components were taken from the native complexed structure, which is somewhat artificial starting point. In the more realistic cases where the individual structures of the constituents were used, these techniques often failed to distinguish the correct orientation (Bacon and Moult 1992). High complementarity of the interacting surfaces in the native complexes is in part due to the "induced fit", e.g. the conformational change in the constituents of the complex upon binding, while individual structures often do not show the perfect matching expected in the complex. There are two general directions in which the simplistic geometric docking algorithms are being improved. First is the introduction of flexibility of ligand and/or receptor to reproduce or mimic the induced fit, and the second is the inclusion of binding determinants other than pure surface complementarity. First attempts to introduce flexibility in protein-protein docking were limited to "softening" of the geometric criteria which would allow certain degree of penetration between the two interacting surfaces (Jiang and Kim 1991, Walls and

Sternberg 1992). Direct simulations with all-atom models may account for the flexibility more accurately and sometimes show promising results (Nilges, M. & Brunger 1993, DiNola *et al.* 1994, Abagyan *et al.* 1994), but are often extremely computationally expensive. Whichever way the flexibility is introduced, it results in much greater ambiguity of the results of geometric docking, since many approximate matches can be found. The multiplicity of solutions calls for additional criteria to select the correct answer. This lead to the inclusion in the docking protocol of the other binding determinants such as estimates of solvation free energy change or molecular mechanics energy (Shoichet and Kuntz 1991), and ultimately, the approximations of the free energy change upon binding (Bohm 1994a,b). Most methods however still use simplistic but faster measures during the generation of the bound conformations and than reevaluate the putative solutions using more sophisticated potentials.

Docking as an energy optimization problem

Complexes considered in the docking studies are in general thermodynamically stable systems. Thus, the native bound conformation should represent the global minimum of the free energy. Consequently, to find the docked conformation, the global minimum of the free energy function of the system has to be located. Since the precise evaluation of the free energy is difficult, one can try to use some approximation that would have similar global minimum. From the energetic point of view, surface complementarity docking methods assume that the interaction energy is proportional to the contact area or other similar measure of the fit of two surfaces, possibly with some penalty for bad contacts (clashes). While this assumption may account reasonably well for van der Waals interactions and, to some extent, for solvation, it obviously disregards the energy contributions from specific pairwise atomic interactions such as hydrogen bond formation and electrostatics. Many recent docking studies try to incorporate these terms, often as

the additional criteria to select the answer from several solutions generated by geometric docking, either using force-field energy evaluation (Shoichet and Kuntz 1991) or elaborate scoring functions (Bohm 1994b, Jain 1996). In several works, physical energy terms were used throughout the algorithm (Abagyan *et al.* 1994, Totrov and Abagyan 1994).

Two major components are required for a successful prediction of the structure of the protein-ligand complex: an efficient global optimization procedure which is capable of finding a global minimum for the strongly anisotropic function of dozens of variables and a free energy approximation for the complex in solution which is computationally inexpensive to be used in the search procedure, yet sufficiently accurate to ensure the uniqueness of the native conformation. In the following two parts we will review the energy calculations and global optimization methods.

Energy terms

Energy calculations are at the center of almost any molecular simulation technique. It is convenient and customary to divide the energy of the molecular system into a number of components, or *energy terms*. Below, five major terms of the molecular interaction energy will be considered in greater detail.

Electrostatic Interactions

Electromagnetism is *the* fundamental force of biochemistry (Davis and McCammon 1990). All processes on the molecular level can be described in terms of electromagnetic interaction combined with quantum mechanical and thermodynamic effects. While covalent and hydrogen bonding as well as Van der Waals interaction all have electrostatic nature, these interactions are complicated by quantum mechanics and it is often convenient to separate them from the longer-

range electrostatic interactions. It is the latter type of interactions which is customarily referred to as electrostatics in biomolecular structure. All proteins and large majority of ligands contain polar atoms interacting strongly with each other and the solvent in the wide range of distances. For a charged amino-acid the strength of electrostatic forces may exceed by more than an order of magnitude the strength of van der Waals interaction (Warshel and Russell 1984).

The evaluation of electrostatic interactions in proteins was first attempted by Lingstrom-Lang in 1924 and a theory of electrostatics in macromolecules was proposed (Tanford and Kirkwood 1957). These macroscopic studies gave some qualitative insights, but only the availability of high-resolution protein structures and computer calculations allowed quantitative studies of protein electrostatics.

The largest problem in electrostatic calculations is the presence of highly polar solvent (water). In vacuum or in the uniform media the interaction between two charges can be simply described by Coulomb law

$$E = k\frac{q_1 q_2}{\varepsilon R_{12}}$$

where $q_{1,2}$ are the charges, R_{12} is the distance between them, ε the dielectric constant and k is 332.0 when the charges are electron units, distance is expressed in angstroms and energy in kcal/mol. In aqueous environment this relation has to be corrected to include the interaction of the charges under consideration with large (virtually infinite) number of surrounding water molecules. Early attempts to simulate macromolecules without consideration of solvent screening ran into difficulties, for example DNA double-helix would be torn apart by electrostatic forces unless the electric charges were drastically reduced (Harvey 1989).

The straightforward and rigorous approach is to include explicitly sufficiently thick layer of water molecules into the calculations. Obviously, it makes calculations heavier, but the principal difficulty of the explicit methods is that liquid water is essentially dynamic environment. Any static placement of water molecules around the system under consideration would result in large errors, as

the physically observed interaction with water is the result of averaging over a large thermodynamic ensemble of the possible states of the solvent. Thus, to achieve accurate results one has to generate this ensemble by an extensive molecular dynamics simulation (Rastelli *et al.* 1995, Simmerling and Elber 1995). While this might be the most rigorous approach to the solvation electrostatic calculations, in most cases it is impractical. Langevine dipoles were proposed (Rossky *et al.* 1978, Luzhkov and Warshel 1992) to make implicit averaging over water molecule's orientations, which eliminates the necessity of generation of a large ensemble of water configurations. The method was applied in protein-protein docking and gave promising results (Jackson *et al.* 1998).

The solvent effectively screens the interaction of the charges of the solute. Generally the farther from each other and the more exposed to the solvent charges are the more their interaction is attenuated. This observation suggested simple corrections to the Coulomb law such as distance-dependant dielectric constant and charge-scaling. While it is somewhat *ad hoc* and doesn't take into account the interaction of the individual charges with the solvent (self-energy), distance-dependant dielectric constant $\varepsilon=\varepsilon_0 R$ is widely used because of it's simplicity (McCammon *et al.* 1979, Pickersgill 1988). This expression actually accelerates calculations of the energy and forces because they become dependent only on R^2 instead of R, eliminating costly square root calculations. Charge scaling was shown to improve the simulation results for such systems as DNA. While these crude approaches can hardly be used for quantitative evaluation of the properties of a macromolecule in solution, they keep the extra calculations to a minimum.

Alternatively, the solvent can be considered as a continuous medium of high dielectric constant. This treatment of solvent is more computationally tractable than the inclusion of explicit water molecules. The electric potential in the medium of variable dielectric constant obeys the Poisson differential equation

$$-\nabla(\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r}))=\rho(\mathbf{r})$$

where ε is the dielectric constant (permittivity), ϕ is the electric potential, and ρ is the charge density. If $\varepsilon(\mathbf{r})=\mathbf{const}$, the Poison equation is equivalent to the Coulomb law, but the solution becomes more complicated when the space is divided into the regions of various dielectric permittivity. Analytic results exist only for special cases such as a sphere. Certain methods utilize these analytic solutions to obtain relatively simple approximations of energy under an assumption that the protein has near-spherical shape, e.g. image method (Friedman 1975, Schaefer and Froemmel 1990, Abagyan and Totrov 1994). Similar assumptions are used in generalized Born approximation (Still *et al.* 1990, Cramer and Truhlar 1992). The precision of these methods is rather limited. Much more rigorous approach is to solve the Poisson equation numerically. Several techniques based on this idea were developed and are widely used in the protein energy calculations (Zauhar and Morgan 1985, Juffer *et al.* 1991, Nicholls and Honig 1991, Zauhar and Varnek 1996). Main difficulty in their application to docking is high computational cost. A hybrid method was recently proposed and used in docking simulation, utilizing single numerical solution of Poisson equation for unbound receptor supplemented by generalized Born–type terms calculated for each specific bound ligand conformation (Majeux *et al.* 1999).

Hydrophobicity

Transfer to the aqueous solution of a number of organic groups results in a free energy loss related to the ordering of water molecules around such groups which is known as hydrophobic effect. The concept of hydrophobic interaction was introduced by Kauzmann (Kauzmann 1959). This effect is similar in nature to the macroscopic surface tension. Hydrophobic interaction is a major driving force in the formation of most ligand-receptor complexes. For some ligands such as steroids the interaction is almost exclusively hydrophobic, and many other ligands are amphiphylic with hydrophobic groups binding into hydrophobic pockets of the

receptor. By fitting the transfer free energies of hydrocarbons against the solvent accessible surface, the hydrophobic contribution was shown (Chothia 1976) to be proportional to the solvent accessible surface with fairly good precision. However, the coefficient of this proportionality is a subject to some controversy since it differs sharply from the microscopically observed value of the surface tension constant. Microscopic surface tension value derived from the transfer energies of aliphatic compounds is close to 30 cal/Å^2 while macroscopic hydrocarbon-water surface tension constant is ~75 cal/Å^2. Some attempts were made to explain the discrepancy by taking into consideration the curvature dependence of the surface tension and the difference of the molar volume of solute and solvent (Sharp *et al.* 1991).

It remains to be seen if the division of the water-solute interaction into solvation electrostatics and hydrophobic components is the most adequate approach. Methods based on this partitioning were shown to reproduce successfully experimental data on transfer free energies for a large set of compounds (Sitkoff at al. 1994). However, alternative approaches to water-solute interaction evaluation were also developed, particularly a number of atomic solvation parameter (ASP) based methods (Eisenberg and McLachlan 1986, Wesson and Eisenberg 1992). ASP methods differentiate the atoms of the solute into a number of types, each with a particular value of solvation energy surface density, generalizing the surface tension. The underlying assumption is that the water-solute interaction can be partitioned into atomic contributions, which are proportional to the solvent accessible surface areas of the atoms. Popularity of ASP approach is in part due to the simplicity and computational efficiency, while the drawbacks are that neither proportionality of the solvation energy to the accessible surface nor the partitioning of the solvation energy into atomic contributions cannot be rigorously justified and are largely *ad hoc* assumptions. Nevertheless, good agreement with experimental data can be achieved (Horton and Lewis 1992), which might in part be explained by the large number of

adjustable parameters in the ASP models. It is questionable that these methods can perform well on a set of compounds which is much larger then the set used for the parameter adjustment.

Van der Waals interactions

The most generic type of interatomic force which exhibits itself as a very strong repulsion at short distances and turns into relatively weak and quickly decreasing attraction as the distance between two atoms grows. It is commonly described by 6-12 potential:

$$E_{VW}(R_{ij}) = -\frac{A_{ij}}{R_{ij}^{6}} + \frac{B_{ij}}{R_{ij}^{12}}$$

where R_{ij} is the distance between the two atoms *i* and *j*. Parameters A_{ij} and B_{ij} depend on the types of atoms and are usually calculated using combination rules from the parameters for the identical pairs of atoms, which are in turn evaluated from quantum-mechanical or experimental data. Usually these parameters are derived along with the other components of the atomic interaction energy to form so-called molecular mechanics force-fields, such as CHARMM (Brooks *et al.* 1983), AMBER (Weiner *et al.* 1984), MMFF (Halgren 1995) and ECEPP (Momany *et al.* 1975). While the $1/R^6$ form of the attraction term has strict quantum-mechanical basis, rigorous description of the repulsion term is more complicated. Alternative forms of the repulsion term have been proposed (e.g. Halgren 1995). Fortunately, the interactions in biomolecular systems occur mostly in the range of inter-atomic distances where attractive term is prevalent, and seem to avoid strong repulsion, alleviating the problem of the exact description of the repulsive term.

Still, extreme sensitivity of the Van der Waals interactions to the small conformational changes makes it's inclusion in the calculation of binding energy problematic. This led a number of authors to simply omit the Van der Waals

contribution in the binding energy, as it seems to introduce more noise than signal into the energy estimates (Krystek *et al.* 1993, Vajda *et al.* 1994). Such omission is partly justified by the cancellation of ligand-receptor interactions in the bound state and the ligand-solvent/receptor-solvent interactions in the unbound state. One can assume that overall number of inter-atomic contacts in the system remains nearly constant upon binding, resulting in the conservation of the total Van der Waals interaction energy. Geometric docking can be used to achieve reasonably good packing. However, this approach leaves out entirely the dependence of the interaction energy on the quality of the interface. In the case of the docking of novel ligands the evaluation of the interface is essential for determination of the binding likelihood and the correct binding mode. Possible compromise is to modify the Van der Waals potential so that it becomes less sensitive to the small deviations in atomic coordinates.

Hydrogen Bonds

Hydrogen bond interaction is a specific attraction between polar hydrogens and a number of heavy atoms (primarily oxygen, nitrogen, sulfur) which have unshared electron pairs. The observations of large number of complexes with solved 3D structures show that many ligands form extensive networks of hydrogen bonds with their receptors, especially in cases of high specificity and high affinity binding. Hydrogen bonds also play important role in the protein folding, where their formation between the turns of the α-helixes and between the β-strands stabilizes these essential secondary structure elements. Unfortunately, there seems to be no agreement so far about the adequate functional form for the hydrogen bonding interaction term and even the energetic value of an average hydrogen bond. Since its origin lays in the same electrostatic and quantum interactions as the origin of Van der Waals and electrostatic terms, hydrogen bonding is often included in the force field as a modification to the Van der Waals potential for the

specific atom pairs (Nemethy *et al.* 1992, Halgren 1995). The modification may only involve change in the parameters (MMFF), or a different functional form (10-12 instead of the standard 6-12 Van der Waals potential in ECEPP). Some force fields simply ignore hydrogen bonding in hope that electrostatic term will provide sufficient favorable contribution when positive hydrogen atoms and negative hydrogen bond acceptors are brought together. However, the charge distribution around the acceptor atoms is highly anisotropic since the unshared electron pairs occupy sp^x orbitals, resulting in strong anisotropy of the HB interaction. High directionality of the HB interaction can also be observed in the solved structures of the proteins and protein complexes (Ippolito *et al.* 1990). This anisotropy is largely ignored by pair-wise, atom-centric potentials used by the majority of the force fields. Such omission may not lead to large errors as long as only naturally occurring conformations are considered since they often already have optimal or sub-optimal configuration of hydrogen bonds. However, in the course of a simulation, such as docking, it may result in erroneous formation of hydrogen bonds of physically impossible geometries. Several forms of hydrogen-bonding term with explicit angular dependence were proposed (Goodford 1985, Miller 1994).

Conformational Entropy

Binding of the ligand to the receptor usually imposes strong constraints upon it's conformational freedom. Also, the surface side-chains of the receptor which are in contact with the ligand may no longer access some of their rotameric states. There is a loss in translational and rotational degrees of freedom, which does not depend on the participating molecules and can be seen as constant as long as only 1-to-1 stoicheometry complexes are considered. Thus, binding may result in substantial decrease in the entropy. As an illustration, one can consider the burial of one CH_2 group in an aliphatic chain. The loss of three rotameric states of the chain results

in the enthropy loss which adds $RT\ln 3 \approx 0.66$ kcal/mole to the free energy of the system, while the decrease in hydrophobic term is around -0.88 kcal/mole (Yang *et al.* 1992).

Exact determination of the entropy change would require extensive molecular dynamics simulations. Currently such simulations are too expensive computationally to use them routinely for a large number of putative complexed structures. Docking methods generally assume that upon binding the ligand is locked in a single conformation. While in some cases this assumption might be far from true, it allows exclusion of the conformational enthropy term from docking simulations as a constant.

Conformational search techniques

An efficient global optimization procedure is a key component of the docking protocol. Many approaches treat both ligand and receptor as rigid bodies (Kuntz *et al.* 1982, Cherfils *et al.* 1991, Bacon and Moult 1992). Such treatment allows for rapid location of the optimal mutual orientation of the two molecules by special techniques (DOCK), but has limited applicability since the majority of small ligands are flexible and structural rearrangements occur in a number of receptors. To some extent, the limitations of the rigid-body docking can be circumvented if several low-energy conformations of the ligand are generated and then docked. The best solution can be than picked as an answer (Kearsley *et al.* 1994, Leach 1994). However, the number of conformations which have to be docked independently to achieve an accurate solution may become very large even for relatively small compounds. Therefore, many techniques try to treat the flexibility of the ligand more directly. Flexible ligand often can be partitioned into rigid fragments. For each fragment, rigid docking can produce a number of favorable orientations. Fragments are then reassembled into the original chemical structure ("Hammerhead" in Welch *et al.* 1996, Gulukota *et al.* 1996). Alternatively, one

fragment is assumed to be essential for binding and placed in the active site first, then others are attached incrementally (Bohm 1992, Rarey *et al.* 1996).

Global optimization of the free energy function with respect to the orientation and the conformation of the ligand is, perhaps, the most strict approach. However, two features of the protein-ligand energy landscape complicate the problem of the energy optimization: high dimensionality and multiplicity of local minima. High dimensionality makes the exhaustive search of the conformational space very computationally expensive. Large number of local minima makes rational determination of the global search direction virtually impossible and limits the usability of the derivatives to a small vicinity of one local minimum. In order to deal with these difficulties, techniques such as Monte-Carlo minimization (Caflisch *et al.* 1992, Totrov and Abagyan 1997, Trosset and Scheraga 1998), Monte-Carlo with simulated annealing (Goodsell and Olson 1990, Goodsell *et al.* 1996), and genetic algorithm (Jones *et al.* 1997) have been applied with various success. Some of these methods use internal coordinates to reduce the dimensionality of the search space.

Monte-Carlo

The term Monte-Carlo has been introduced by Metropolis and Ulam (Metropolis *et al.* 1953), with an allusion to the essentially random nature of such simulations. Monte-Carlo minimization consists of three repetitive steps:

1. Random Jump. One or several variables in the system are changed randomly.

2. Local Minimization. The energy of randomized conformation is optimized using conjugate gradient or quasi-Newton technique to achieve a new local minimum.

3. Evaluation. New conformation is accepted or rejected according to the Metropolis criterion: if the energy of the new conformation E_{new} is lower than the

energy of the old one E_{old} , new conformation is always accepted and used in the next iteration. Otherwise, it is accepted with the probability of

$P_{acc}=\exp(-(E_{new}-E_{old})/kT)$, where k is Boltzman's constant and T is the effective temperature of the simulation.

It has been established that a full local minimization after each random step greatly improves the efficiency of the procedure (Li and Scheraga 1987, Abagyan and Argos 1992). However, some components of the energy, such as solvation electrostatic energy, might have no derivatives and/or might be too computationally expensive for local minimization. Double-energy MC minimization scheme (Abagyan *et al.* 1994) circumvents this obstacle by using two sets of energy terms, one for the local gradient minimization stage and another one for the Metropolis criterion evaluation stage in the MC step. Such division can be justified if the extra terms included for the Metropolis criterion are relatively "slow", insensitive to small conformational changes.

Internal coordinates

One of the principal difficulties in biomolecular simulations is the size of the system which often contains thousands of atoms. As a consequence, the conformational space has a very high dimensionality, complicating the search for the global energy minimum. The use of internal coordinates substantially reduces the number of variables defining the conformation of the system. Cartesian description requires 3 variables (x,y,z) per atom . Internal coordinates description uses bond lengths, planar angles and torsion angles instead. Since bond lengths and planar angles are essentially rigid at normal conditions, one can consider them as constants and only allow torsion angle changes (rotations around the bonds), reducing the dimensionality of conformational space at least threefold. Practically even greater reduction is achieved since at every branching point several atoms share the same torsion angle (Fig 1).

The formal geometrical description to allow efficient manipulations of the multi-molecular system in internal coordinates with arbitrary subsets of free and fixed variables was introduced (Abagyan *et al.* 1994). The technique represents the system as a directed treelike graph imposed on all atoms as well as on some auxiliary virtual atoms (Fig 1). Each atom in this basic description has three geometric parameters determining its position with respect to the preceding part of the tree. The parameters are bond length b, bond angle ω and torsion φ or phase ϕ dihedral angles for the main branch and side branches, respectively. The sub-trees of different molecules join in the starting triple of virtual atoms which are fixed at the origin of the coordinate system and allow for standard treatment of all real atoms including the root atoms of each molecular sub-tree. When several internal variables are fixed (considered constant) a group of atoms may form so-called rigid-body, where mutual positions of the atoms involved do not change upon any changes of the remaining free variables. The concept of rigid bodies provides an important additional advantage for the energy calculations, since all pair-wise energy contributions from the atoms within a rigid body are constant. Such contributions often can be excluded from the calculations when only the relative energy change is important, improving the computational performance.

Other approaches

Various global optimization techniques were applied to the docking problem. Among the more popular is the genetic algorithm (GA), which was widely applied in protein folding simulations (Clearwater 1991, Unger and Moult 1993, Dandekar and Argos 1994). The idea of GA is to mimic the evolution process by manipulating "chromosomes", each containing a set of variable values defining a possible solution, e.g. a certain binding mode. The values inside the "chromosome" might be the rotatable torsion angles of the ligand and the variables defining the relative orientation of the ligand and receptor. The algorithm starts

with a random "population" of chromosomes, from which new generations are produced by "mutations" and "crossovers", which involve, respectively, randomization of some variables inside the chromosome or reshuffling of some variable values between two chromosomes. The best-fit "individuals" are preserved while others are discarded according to the fitness function. The assumption is that as the algorithm progresses, this strategy will find and keep the advantageous combinations of variable values, converging to the minimum of the fitness function. The GA docking was used fairly successfully to reconstitute a large number of known complexes (Jones *et al.* 1997), although no tests were undertaken to compare its performance with more conventional approaches such as MC.

Notably, Fourier-transform was also used to locate the optimal geometric fit (Katchalski-Katzir *et al.* 1992). The method is efficient and attractively simple conceptually. Unfortunately it seems to be only be applicable to a rather simplistic fitness function and can only optimize efficiently the three translational degrees of freedom. Rotations still need to be sampled by other means, i.e. systematic or random search. Fourier-transform approach may be useful primarily in the cases where the interacting molecules are very big, making other methods too expensive computationally.

Molecular dynamics (MD) simulation can be used as an optimization method, and potentially it can provide a realistic picture of the binding process. However, MD is the most computationally expensive approach, and so far it is impossible to simulate the whole progress of the system from unbound components to the complex. The use of MD in docking is now limited to the simulations of the already bound complexes, where it is successfully used to predict various thermodynamic properties (Rosenfeld *et al.* 1995, Miranker and Karplus 1991, DiNola *et al.* 1994, Luty *et al.* 1995). Somewhat better performance can be achieved using so-called Brownian dynamics (Rossky *et al.* 1978), which was applied to simulate

long-range diffusion-like motions of the interacting macromolecules (Kozack and Subramaniam 1993).

An example docking study on a set of protein-ligand complexes with known 3D structures.

We tested the ability of internal coordinate MC minimization docking procedure to predict the native conformations of protein-ligand complexes using a benchmark set of 51 high-resolution structures from PDB. Ligands were diverse in size, from 12 to 84 atoms, and had a broad range of chemical properties and included sugars, fatty acids, phosphates, bases, heterocyclic and other compounds, which insured the applicability of the docking procedure to a large variety of receptor/ligand pairs.

Methods

Energy

Our energy estimate used during the docking simulations consisted of the following terms:

$$E = E_{FFint} + E_{VW} + E_{HB} + E_{HP} + E_{EL}$$

ΔE_{FFint} is the force-field energy which included internal Van der Waals interactions and torsion energy for the ligand calculated with ECEPP/3 parameters (Nemethy *et al.* 1992). Since ECEPP/3 only has parameters for amino-acid atom types, the atoms of ligands were assigned closest chemically similar atom types. The rest of the terms refer to inter-molecular interactions.

Because of its extreme rigidity, Van der Waals potential in its standard 6-12 form may introduce large noise in the energy function. For inter-molecular interactions we therefore used a modified smoother form of the potential with

most of the repulsive part truncated. Truncation was achieved by the following transformation of the original value of Van der Waals potential :

$$E_{VW} = \begin{cases} E^0_{VW}, if & E^0_{VW} \le 0 \\ \dfrac{E^0_{VW} E_{\max}}{E^0_{VW} + E_{\max}}, if & E^0_{VW} > 0 \end{cases}$$

This expression ensures smooth transition from undistorted form of Van der Waals potential in the negative range of values to increasingly attenuated form in the positive range, asymptotically approaching E_{max} cutoff value. E_{max} was chosen on the basis of preliminary tests to be 1.5 kcal/mole. Lower values sometimes result in severely clashed docking solutions as the Van der Waals repulsion is no longer able to compete with attractive terms, primarily electrostatics. This and other interaction potentials were precalculated on a grid to accelerate energy evaluation during the simulations. The grid cell size was set to 0.5Å.

ΔE_{HB} is hydrogen bonding term which was calculated using Gaussian-type potential positioned around the center of each lone electron pair of the hydrogen-bond acceptors:

$$E_{HB} = E^0_{HB} e^{\frac{(r-r_p)^2}{d^2_{HB}}}$$

The peak interaction energy E^0_{HB} was assumed to be 2.5 kcal/mol as an average of various estimates, and the radius of the interaction sphere d_{HB} was assumed to be 1.4Å, allowing for about 30° to 40° deviation from the ideal geometry in accordance with observations in X-ray structures. $\mathbf{r}_{hb}$ is the radius-vector of the interaction center, which was placed 1.7Å from the atom. In case of hydrogen atoms the center was placed along the axis of the covalent bond attaching the hydrogen to the rest of the molecule. In case of heavy sp^2 atoms, one (for nitrogen) or two (for oxygen) centers were placed at the angle of 120° to the existing covalent bond. For sp^3 oxygen and sulfur, two centers were placed in tetrahedral geometry, at 109° to the existing covalent bonds and to each other.

Electrostatic term E_{EL} used modified Coulomb law with distance dependant dielectric constant $\varepsilon=4r$. Hydrophobic term E_{HP} was calculated as roughly proportional to the buried hydrophobic surface with the free energy density of 30 cal/mol/Å^2. To accelerate calculations, a grid-based form of the hydrophobic potential was developed. The fragments of the solvent-accessible surface were generated using the modified Shrake and Rupley algorithm (Shrake and Rupley 1973, Abagyan *et al.* 1994). The algorithm produces dots which evenly cover the surface. The hydrophobic potential on the grid was then calculated as:

$$E_{HP} = E^0_{HP} e^{-\frac{d^2_{surf}}{d^2_W}}$$

d_{surf} is the distance to the closest point of the hydrophobic surface, and d_w is effective radius of the hydrophobic interaction which was set to the diameter of the water molecule 2.8Å. The value of E^0_{HP}=3 kcal/mole was chosen to approximate the surface tension of 30 cal/mol/Å^2 for extended hydrophobic surfaces in test cases.

Conformational search procedure

All ligand /receptor pairs in the set were docked using flexible Monte-Carlo docking procedure with potential maps as implemented in ICM software (Totrov and Abagyan 1997, Abagyan *et al.* 1994, Abagyan and Totrov 1994). The ICM method describes both the relative positions of two molecules and their conformations by a uniform set of internal variables. Any subset of internal variables can be subjected to local or global energy minimization procedures. In this study, the global Monte-Carlo minimization procedure similar to previously described (Totrov and Abagyan 1997, Abagyan *et al.* 1994) was used. It involved random conformational change of two possible types: positional Pseudo-Brownian random move or internal torsion modification, followed by local energy minimization (up to 100 steps of conjugate gradient minimization) and selection by the Metropolis criterion (temperature factor was set to 600K). Pseudo-

Brownian random moves changed the position of the ligand molecule as a whole with a certain amplitude (here we used 2Å), as well as randomly rotated it around it's center of gravity by an angle close to the translation amplitude over the radius of gyration. Internal torsion angles of the ligand were randomly changed one at a time, with the amplitude of 180°.

Geometrically different (as evaluated by the root mean square displacement of the ligand atoms) and low energy conformations were accumulated in the conformational stack (Abagyan and Argos 1992). Adaptive length of the MC runs was used, with the limit on the total number of steps proportional to the size (number of atoms) of the ligand: $N_{MCsteps}=50*N_{LigAtom}$. Similarly, an adaptive length of local minimization during the MC run was used: $N_{LocMinSteps}=25+N_{LigAtom}$. The factors in these relations were established empirically from the convergence and efficiency considerations.

Test data set

The set of 51 complexes (Table 1.) was extracted from high-resolution PDB structures. The structures were selected according to a number of criteria: We discarded all structures at resolutions worse than 2.0Å since large errors in the receptor coordinates could result in poor docking and recognition for reasons unrelated to our study. Some complexes had the ligand bound covalently to the receptor and were also discarded since the prediction of such chemical reactions is beyond the scope of our approach. We also omitted complexes where metal ions were directly involved in the protein-ligand interaction since the force field used in the simulations did not provide for adequate modeling of such atoms. For a number of receptors structures of several complexes with different ligands were available. In such cases we used a single receptor structure in docking experiments with all ligands. Hydrogen atoms were added to all X-ray structures using the hydrogen placement algorithm of ICM software (Abagyan *et al.* 1994). Electric

charges were assigned to the atoms of the ligands using bond-charge increment algorithm from MMFF94 force field (Halgren 1995).

Results and discussion

51 complexes with known structures were predicted. Only the best-energy conformation in each case was retained and compared to the experimental structure. 35 predictions were within 3Å from the native structure, producing correct overall positioning of the ligand, and 26 were within 2Å, giving fairly detailed picture of the receptor-ligand interaction (Table 1, Fig.2). As expected, good precision is achieved for tighter, enclosed binding pockets, while for more loose, open binding sites such as in phospholipase, FK506 binding protein or fatty acid binding protein the prediction quality is often marginal. Single simulation took from 2 to 12 min CPU time, which illustrates the advantage of pre-calculated grid potentials, since similar simulations with full-atom receptor molecule take several hours (Totrov and Abagyan 1997).

The results show that docking techniques, such as flexible docking in internal coordinates using grid potential representation of the receptor molecule, in the majority of cases can produce a model of protein-ligand interaction with the precision allowing its use in applications such as drug design. However, an important condition for the current docking methods is the relative rigidity of the binding site. Reliable ways to treat receptor flexibility are yet to be developed. The growth in the available computer power and improvement in simulation techniques should ultimately allow detailed predictions of flexible receptor and ligand interaction.

References

Abagyan, R.A. & Argos, P. (1992). Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *J. Mol. Biol.*, **225**, 519-532.

Abagyan, R.A. & Totrov, M.M. (1994). Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins. *J.Mol.Biol.*, **235**, 983-1002.

Abagyan, R.A., Totrov, M.M. & Kuznetsov, D.A. (1994). ICM: a new method for structure modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comp. Chem.*, **15**, 488-506.

Bacon, D.J. & Moult, J. (1992). Docking by Least-squares Fitting of Molecular Surface Patterns. *J. Mol. Biol.*, **225**, 849-858.

Beddell CR, Goodford PJ, Norrington FE, Wilkinson S, Wootton R (1976). Compounds designed to fit a site of known structure in human haemoglobin. *Br J Pharmacol*, **57(2)**, 201-209.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535-542.

Bohm, H.J. (1992). LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aided Mol. Design*, **6**, 593-606.

Bohm, H.J. (1994). On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three-dimensional structure. *J Comput Aided Mol Des*, **8(5)**, 623-632.

Bohm, H.J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des*, **8(3)**, 243-256.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J. Swaminathan, S. & Karplus, M. (1983). CHARMM: A Program for Macromolecular Energy, Minimization and Dynamics Calculations. *J. Comput. Chem.*, **4**, 187-217.

Caflisch, A., Niederer, P., Anliker, M. (1992). Monte Carlo docking of oligopeptides to proteins. *Proteins: Struct., Funct. & Gen.*, **13**, 223-230.

Cherfils, J., Duquerroy, S. & Janin, J. (1991). Protein-protein recognition analyzed by docking simulation. *Proteins: Struct., Funct. & Gen.*, **11**, 271-280.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.*, **105**, 1-12.

Clearwater, S.H., Huberman, B.A., Hogg, T. (1991). Cooperative solution of constraint satisfaction problems. *Science*, **254**, 1181-1183.

Cramer, C.J. & Truhlar, D.G. (1992). An SCF solvation model for the hydrophobic effect and absolute free energies of aqueous solvation. *Nature*, **256**, 213-217.

Dandekar, T. and Argos, P. (1994). Folding the main chain of small proteins with the genetic algorithm. *J. Mol. Bio.*, **236**, 844-861.

Davis, M.E. & McCammon, J.A. (1990). Electrostatics in Biomolecular Structure and Dynamics. *Chem. Rev.*, **90**, 509-521.

DesJarlais RL, Sheridan RP, Dixon JS, Kuntz ID, Venkataraghavan R (1986). Docking flexible ligands to macromolecular receptors by molecular shape *J Med Chem*, **29**, 2149-2153.

DiNola, A., Raccatano, D. & Berendsen, H. (1994). Molecular dynamics simulation of the docking of substrates to proteins. *Proteins*, **19**, 174-182.

Eisenberg, D. & McLachlan, A.D. (1986). Solvation Energy in Protein Folding and Binding. *Nature*, **316**, 199-203.

Fischer, D., Lin, S.L., Wolfson, H.L., Nussinov, R (1995). A geometry-based suite of molecular docking processes. *J. Mol. Biol.*, **248**, 459-477.

Friedman, H.L. (1975). Image approximation to the reaction field. *Molecular Physics*, **29**, 1533-1543.

Goodford, P.J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, **28**, 849-875.

Goodsell A.S.; Olson A.J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct., Funct. & Gen.*, **8**, 195-202.

Goodsell, D.S., Morris, G.M., Olson, A.J (1996). Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recognit.*, **9**, 1-5.

Gulukota, K., Vajda, S. & Delisi, C. (1996). Peptide Docking Using Dynamic Programming. *J. Comp. Chem.*, **17**, 418-428.

Halgren, T.A. (1995). Merck Molecular Force Field. I.-V. *J. Comp. Chem.*, **17**, 490-641.

Harvey, S. (1989). Treatment of Electrostatic Effects in Macromolecular Modeling *Proteins: Struct., Funct. & Gen.*, **5**, 78-92.

Helmer-Citterich M, Tramontano A (1994). PUZZLE: a new method for automated protein docking based on surface shape complementarity. *J Mol Biol.*, **235(3)**, 1021-1031.

Horton N, Lewis M (1992). Calculation of the free energy of association for protein complexes. *Protein Sci.*, **1(1)**, 169-181.

Ippolito JA, Alexander RS, Christianson DW (1990). Hydrogen bond stereochemistry in protein structure and function. *J Mol Biol* , **215(3)**, 457-471.

Jackson RM, Gabb HA, Sternberg MJ (1998). Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J Mol Biol*, **276(1)**, 265-285.

Jain AN (1996). Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J Comput Aided Mol Des*, **10(5)**, 427-440.

Jiang, F. & Kim S.-H. (1991). Soft docking *J. Mol. Biol.*, **219**, 79-102.

G. Jones, P. Willett, R.C. Glen, A.R. Leach, and R. Taylor (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, **267**, 727-748.

Juffer, A.H., Botta, E.F.F., van Keulen, B.A.M., van der Ploeg, A. & Berendsen, H.J.C. (1991). The electric potential of a macromolecule in a solvent: a fundamental approach. *J. Comput. Phys.*, **97**, 144-171.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., Vakser, I.A (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A*, **89**, 2195-9.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, **14**, 1-63.

Kearsley, S.K., Underwood, D.J., Sheridan, R.P., Miller, M.D. (1994). Flexibases: a way to enhance the use of molecular docking methods. *J Comput Aided Mol. Design*, **8**, 565-82.

Kozack RE, Subramaniam S (1993). Brownian dynamics simulations of molecular recognition in an antibody-antigen system. *Protein Sci.*, **2(6)**, 915-926.

Krystek S, Stouch T, Novotny J (1993). Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures. *J Mol Biol*, **234(3)**, 661-679.

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. & Ferrin, T.E. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, **161**, 269-288.

Leach, A.R. & Kuntz, I.D. (1992). Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comp. Chem.*, **13**, 733-748.

Leach, A.R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.*, **235**, 345-356.

Li, Z. & Scheraga, H.A. (1987). Monte Carlo-Minimization Approach to the Multiple-Minima Problem in Protein Folding. *Proc. Natl. Acad. Sci. USA*, **84**, 6611-6615.

Luty, B.A., Wasserman, Z.R., Stouten, P.F.W., Hodge, C.N., Zacharias, M., McCammon, J.A. (1995). A molecular mechanics/grid method for evaluation of ligand-receptor interactions. *J. Comput. Chem.*, **16**, 454-464.

Luzhkov V, Warshel A (1992). Microscopic models for quantum mechanical calculations of chemical processes in solution: LD/AMPAC and SCAAS/AMPAC calculations of solvation energies. *J. Comp. Chem.*, **13**, 199-213.

Majeux N, Scarsi M, Apostolakis J, Ehrhardt C, Caflisch A (1999). Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins*, **37(1)**, 88-105.

McCammon, J.A., Wolynes, P.G. & Karplus, M. (1979). Picosecond Dynamics of Tyrosine Side Chains in Proteins. *Biochemistry*, **18**, 927-942.

Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, N.M., Teller, A.H., Teller, E. (1953). Equation of State calculations by Fast Computing Machines. *J. Chem. Phys.*, **21**, 1087-1092.

Miller MD, Kearsley SK, Underwood DJ, Sheridan RP (1994). FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J Comput Aided Mol Des*, **8(2)**, 153-174.

Miranker, A. & Karplus, M. (1991). Functionality maps of binding cites: a multiple copy simultaneous search method. *Proteins: Struct., Funct. & Dyn.*, **11**, 29-34.

Momany, F.A., McGuire, R.F., Burgess, A.W. & Scheraga, H.A. (1975). Energy Parameters in Polypeptides. VII. Geometric Parameters, Partial Atomic Charges, Nonbonded Interactions, Hydrogen Bond Interactions, and Intrinsic Torsional Potentials for the Naturally Occurring Amino Acids. *J. Phys. Chem.*, **79**, 2361-2381.

Nemethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S. & Scheraga, H.A. (1992). Energy Parameters in Polypeptides. 10. Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.*, **96**, 6472-6484.

Nicholls, A. & Honig, B. (1991). A Rapid Finite Difference Algorithm, Utilizing Successive over-Relaxation to Solve the Poisson-Boltzmann Equation. *J. Comput. Chem.*, **12**, 435-445.

Nilges, M. & Brunger, A. (1993). Successful Prediction of the Coiled Coil Geometry of the GCN4 Leucine Zipper Domain by Simulated Annealing: Comparison to the X-Ray Structure. *Proteins*, **15**, 133-146.

Pickersgill, R.W. (1988). A rapid method of calculating charge-charge interaction energies in proteins. *Prot. Eng.*, **2**, 247-248.

Rarey, M., Kramer, B., Lengauer, T., Klebe, G (1996). A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, **261**, 470-489.

Rastelli G., Thomas B., Kollman P.A., Santi D.V. (1995). Insight into the specificity of thymidilate synthase from molecular dynamics and free energy perturbation calculations *J Am Chem Soc*, **117**, 7213-7227.

Rossky, P.J., Doll, J.D. & Friedman, H.L. (1978). Brownian Dynamics as Smart Monte Carlo Simulation. *J. Chem. Phys.*, **69**, 4628-4633.

Rosenfeld, R., Vajda, S. & DeLisi, C. (1995). Flexible docking and design. *Annu. Rev. Biophys. Biomol. Struct.*, **24**, 677-700.

Schaefer, M. & Froemmel, C. (1990). A Precise Analytical Method for Calculating the Electrostatic Energy of Macromolecules in Aqueous Solution. *J. Mol. Biol.*, **216**, 1045-1066.

Sharp, K.A., Nicholls, A., Fine, R.F. & Honig, B. (1991). Reconciling the Magnitude of the Microscopic and Macroscopic Hydrophobic Effects. *Science*, **252**, 106-109.

Shoichet, B.K. & Kuntz, I.D. (1991). Protein Docking and Complementarity. *J. Mol. Biol.*, **221**, 327-346.

Shrake, A. & Rupley, J.A. (1973). Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin. *J. Mol. Biol.*, **79**, 351-371.

Simmerling, C.L., Elber, R. (1995). Computer determination of peptide conformations in water: different roads to structure. *Proc Natl Acad Sci USA*, **92**, 3190-3193.

Sitkoff, D., Sharp, K.A., and Honig B. (1994). Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *Journal of Physical Chemistry*, **98** (7), 1978-1988.

Still, W.C., Tempczyk, A., Hawley, R.C. & Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, **112**, 6127-6129.

Tanford, C. & Kirkwood, J.G. (1957). Theory of Protein Titration Curves. I. General Equations for Impenetrable Spheres. *J. Amer. Chem. Soc.*, **79**, 5333-5339.

Totrov, M.. & Abagyan, R. (1994). Detailed ab initio prediction of lysozyme-antibody complex with 1.6A accuracy. *Nature Structural Biology*, **1**, 259-263.

Totrov, M. & Abagyan, R., (1997). Flexible protien-ligand docking by global energy optimization in internal coordinates *Proteins*, Suppl. 1, 215-220.

Trosset JY, Scheraga HA (1998). Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines. *Proc Natl Acad Sci USA*, **95(14)**, 8011-8015.

Unger, R. and Moult, J. (1993). Genetic algorithms for protein folding simulations. *J. Mol. Biol.*, **231**, 75-81.

Vajda S, Weng Z, Rosenfeld R, DeLisi C (1994). Effect of conformational flexibility and solvation on receptor-ligand binding free energies. *Biochemistry*, **33(47)**, 13977-13988.

Walls, P.H. & Sternberg, M.J.E. (1992). New algorithm to model protein-protein recognition based on surface complementarity. *J. Mol. Biol.*, **228**, 277-297.

Warshel, A. & Russell, S.T. (1984). Calculation of electrostatic interactions in biological systems and in solution. *Quart.Rev.Biophys.*, **17**, 283-422.

Weiner, S.J., Kollman, P.A., Case, D.A., Chandra Singth, U., Ghio, C., Alagona, G., Prefeta, S., Jr. & Wiener, P. (1984). A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Amer. Chem. Soc.*, **106**, 765-783.

Welch, W., Ruppert, J., Jain, A.J. (1996). Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chemistry & Biology*, **3**, 449-462.

Wesson, L, and Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.*, **1**, 227-235.

Yang, A.-S., Sharp, K.A. & Honig, B. (1992). Analysis of the Heat Capacity Dependence of Protein Folding. *J. Mol. Biol.*, **227**, 889-900.

Zauhar, R.J. and Morgan, R.S. (1985). A new method for computing the macromolecular electric potential. *J.Mol.Biol.*, **186**, 815-820.

Zauhar, R.J. and Varnek, A. (1996). A fast and Space-Efficient boundary element method for computing electrostatic and hydration effects in large molecules. *J. Comp. Chem.*, **17**, 864-877.

Table 1 Ligands/receptor pairs used in docking simulations

Source (PDB code)	Receptor	Ligand	RMSD Å
1hmr[1]	Fatty acid binding protein	elaidic acid	4.89
1hms	Fatty acid binding protein	oleic acid	6.69
1hmt	Fatty acid binding protein	stearic acid	3.49
1icm[1]	Intestinal fatty acid binding protein	Myristate	1.79
1icn	Intestinal fatty acid binding protein	oleic acid	1.46
4dfr[1]	DHFR	Methotrexate	1.81
1dyj	DHFR	5,10-dideazatetrahydrofolate	2.84
1dyh	DHFR	5-deazafolate	2.48
1dyi	DHFR	Folate	3.45
1jom	DHFR	folinic acid	5.00
2tbs[1]	Trypsin	Benzamidine	1.91
1tng	Trypsin	Aminomethylcyclohexane	1.27
1tnh	Trypsin	4-fluorobenzylamine	1.94
1tni	Trypsin	4-phenylbutylamine	3.05
1tnj	Trypsin	2-phenylethylamine	2.79
1tnk	Trypsin	3-phenylpropylamine	2.60
1tnl	Trypsin	Tranylcypromine	2.11
188l[1]	Lysozyme mutant	o-xylene	0.42
185l	Lysozyme mutant	Indole	1.33
184l	Lysozyme mutant	Isobutylbenzene	4.53
187l	Lysozyme mutant	p-xylene	1.75
186l	Lysozyme mutant	n-butylbenzene	1.62
181l	Lysozyme mutant	Benzene	0.80
183l	Lysozyme mutant	Indene	0.52
182l	Lysozyme mutant	Benzofuran	0.43
1erb[1]	Retinol binding protein	n-ethyl retinamide	0.99
1fel	Retinol binding protein	Fenretinide	2.21
1fem	Retinol binding protein	Retinoic acid	2.35
1fen	Retinol binding protein	Axerophthene	0.93
1sre[1]	Streptavidin	Haba	1.50

1srg	Streptavidin	3'-methyl-haba	7.24
1sri	Streptavidin	3',5'-dimethyl-haba	7.53
1srj	Streptavidin	Naphthyl-haba	0.76
1gar	Glycinamide ribonucleotide transformylase	Burroughs-Wellcome inhibitor 1476u89	2.19
1fnd	Ferredoxin reductase	FAD	8.87[2]
1ake	Adenylate kinase	Inhibitor ap5a	0.98
1rcf	Flavodoxin	flavin mononucleotide	1.2
1mrj	α-momorcharin	Adenine	3.51
1mrg	α-trichosanthin	Adenine	0.42
1mdq	Maltodextrin-binding protein	Maltose	0.92
1gca	Glucose/galactose-binding protein	Galactose	1.27
2dri	D-ribose-binding protein	beta-d-ribose	0.56
1lst	Lysine-, arginine-, ornithine-binding protein	Lysine	0.61
1hsl	Histidine-binding protein	Histidine	1.68
1ars	Aspartate aminotransferase	Pyridoxal-5'-phosphate	2.37
1fkh	FK506 binding protei	(1r)-1-cyclohexyl-3-phenyl-1-propyl (2s)-1-(3,3-dimethyl- 1,2-dioxopentyl)-2-piperidinecarboxylate	2.27
1mai	Phospholipase c δ-*1*	Inositol trisphosphate	5.03
1nsc[1]	Neuraminidase	sialic acid	0.93
1nsd	Neuraminidase	2,3-dehydro-2-deoxy-n-acetyl neuraminic acid	0.75
1ivd	Neuraminidase	4-(acetylamino)-3-hydroxy-5-nitrobenzoic acid	0.83
1fgi	FGF receptor kinase	Inhibitor SU5402	0.76

1. Structure of the receptor from this PDB entry was used in all docking simulations involving the same protein.
2. Flavine nucleotide moiety was docked well, while the other half (adenosine) deviates significantly from its native position.

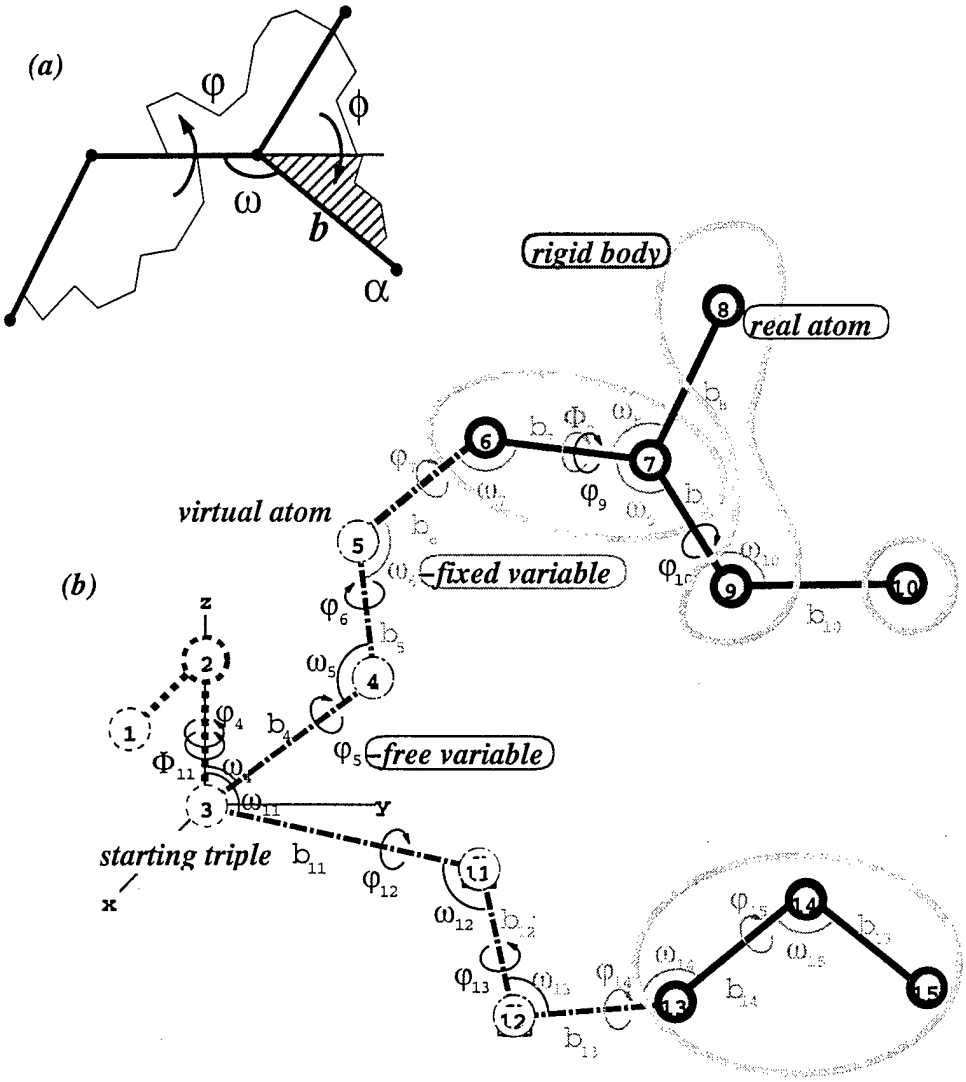
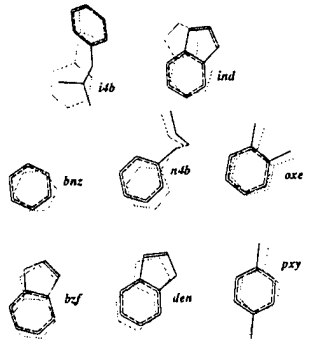

(a)
φ
ϕ
ω
b
α
rigid body
real atom
virtual atom
fixed variable
free variable
(b)
z
y
x
starting triple

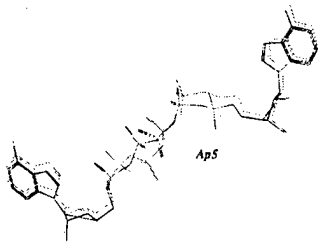
Fig. 1(a) Four types of internal variables considered in ICM. (b) The ICM tree representing the geometry of multimolecular arbitrarily fixed system and containing both real atoms and bonds (continuous lines) and virtual ones (dot-dashed lines). Atoms are numbered so that any atom in the directed graph starts a subtree with a continuous numeration. An arbitrary subset of free internal variables is shown in bold black characters, all the others being fixed (gray characters). The atomic regular directed graph is the basic one, the orderof variables and rigid bodies following it. The numeration does not change as a result of refixation and redefinition of the rigid bodies. The attribution of the main (torsion) branch at the branching point is arbitrary and does not necessarily follow the atomic numeration.

Fig. 2(a,b) Comparison of predictions and experimentally determined structures for 53 protein-ligand complexes.

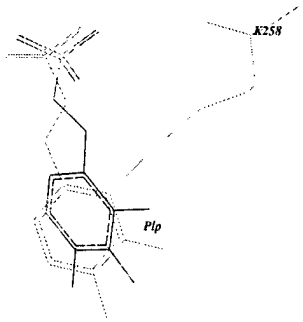
Lysozyme mutant complexes with various small aromatic and hetroaromatic compounds



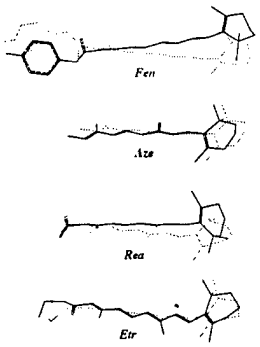
Adenylate kinase complex with with the inhibitor ap5a



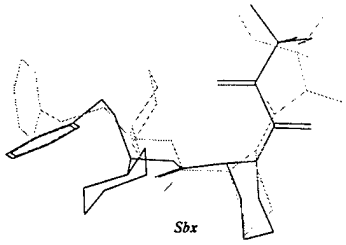
Aspartate aminotransferase complexed with pyridoxal-5'-phosphate



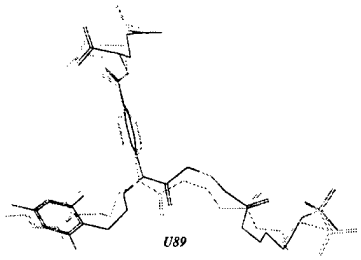
Retinol binding protein complexes with n-ethyl retinamide, fenretinide, retinoic acid and axerophthene



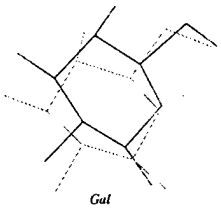
FK506 binding protein complex with (1r)-1-cyclohexyl-3-phenyl-1-propyl(2s)-1-(3,3-dimethyl- 1,2-dioxopentyl)-2-piperidinecarboxylate



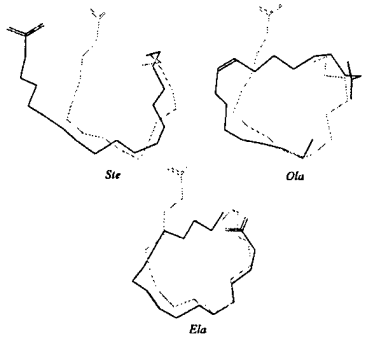
Glycinamide ribonucleotide transformylase complex with Burroughs-Wellcome inhibitor 1476u89



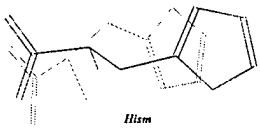
Glucose/galactose-binding protein complex with galactose



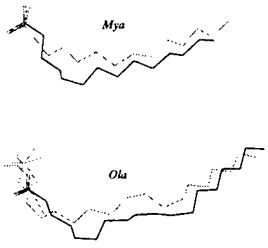
Fatty acid binding protein complexes with elaidic, oleic and stearic acids



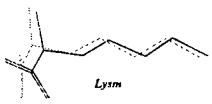
Histidine-binding protein complex with histidine



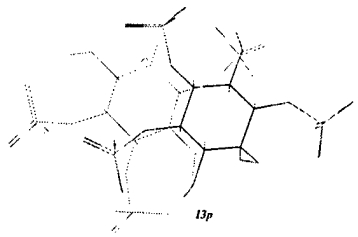
Intestinal fatty acid binding protein complexes with myristate and oleate



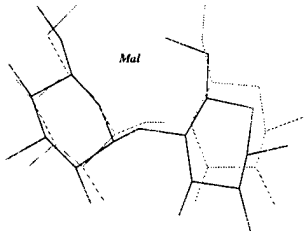
Lysine-, arginine-, ornithine-binding protein complex with lysine



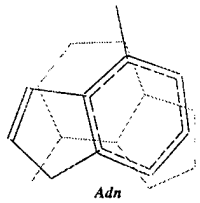
Phospholipase c d-1 complex with inositol trisphosphate



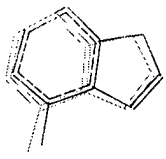
Maltodextrin-binding protein complex with maltose



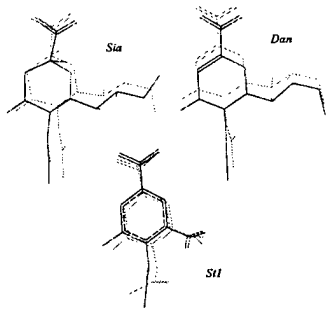
α-momorcharin complex with adenine



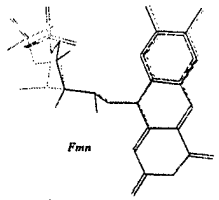
a-trichosanthin complex with adenine



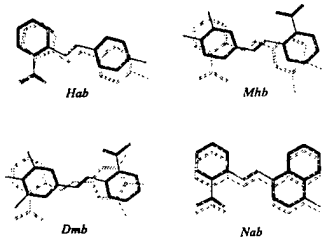
Neuraminidase complexes with sialic acid,
2,3-dehydro-2-deoxy-n-acetyl neuraminic acid
and 4-(acetylamino)-3-hydroxy-5-nitrobenzoic acid



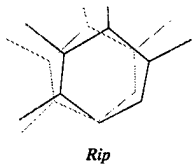
Flavodoxin complex with flavin mononucleotide



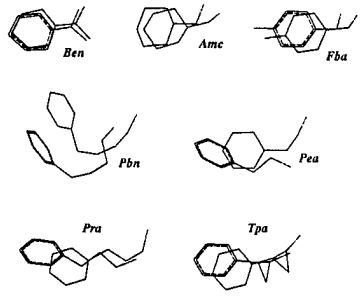
Streptavidin complexes with
2-((4'-hydroxyphenil)-azo)benzoate (HABA),
3'-methyl-HABA, 3',5'-dimethyl-HABA and naphthyl-HABA



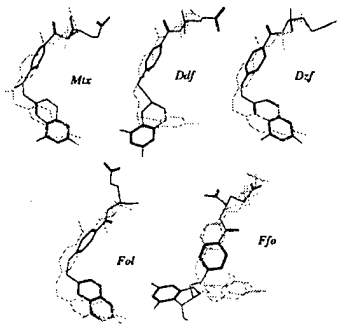
D-ribose-binding protein complex with beta-d-ribose



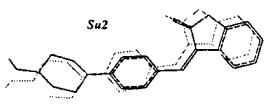
Trypsin complexes with inhibitors benzamidine
aminomethylcyclohexane, 4-fluorobenzylamine 4-phenylbutylamine
2-phenylethylamine 3-phenylpropylamine tranylcypromine



Dihydrofolate reductase complexes with methotrexate,
5,10-dideazatetrahydrofolate, 5-deazafolate, folate and folinic acid



Tyrosine kinase of FGF receptor complex with Sugen inhibitor



Comparative Study of Several Algorithms for Flexible Docking

Badry D. Bursulaya,[1] Maxim Totrov,[1] Ruben Abagyan[1,2] and Charles L. Brooks III[1]

[1]*Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037*

[2]*Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, CA 92121*

Abstract

We have tested several programs for flexible molecular docking: DOCK 4.0, FlexX 1.8, AutoDock 3.0 and ICM 2.8. This was done by doing two kinds of studies: docking experiments on a data set of 37 protein–ligand complexes and screening a library containing 10,037 entries against 11 different proteins. The docking accuracy of the methods was judged based on the corresponding rank 1 solutions. We found that on average the solutions produced by ICM have the highest accuracy. All other methods are less accurate and their predictions are of approximately the same accuracy. The efficiency of docking, which measure a computational cost of achieved accuracy, is very low in case of AutoDock and approximately the same for all other methods. The database screening was performed using DOCK, FlexX and ICM. In 19 cases ICM is able to find original ligands within the top 1 % of the total library. The corresponding number for DOCK and FlexX is 8 and 11 respectively.

1 Introduction

Molecular docking has become a useful tool in drug discovery efforts. The screening of large databases for possible lead compounds is nowdays becoming a routine procedure. Until recently the molecular docking was performed using a single conformation for each ligand. This approach, which treats ligand as a rigid body, is CPU benign and thus satisfies a key requirement that docking algorithm should not spend more than few minutes per compound while searching the database. However, the constraints introduced by fixing the internal degrees of freedom of the ligand, although advantageous in terms of computational cost, could have a negative impact on the ability to make a valuable prediction. In particular, the freezing of internal motion may prevent ligand from adopting a conformation that would fit better into a binding site of a receptor. Thus efforts has been advanced to develop algorithms for a flexible ligand docking. Based on the approach to the conformational search of flexible ligand, we can group the algorithms in the following two major classes: algorithms, which try to fit the ligand into the binding pocket of the protein by matching (geometrically, chemically, energetically etc.); and algorithms, which find an optimum ligand conformation by solving a global energy optimization problem. In this study we attempt to evaluate the performance of several programs employing algorithms of both classes. The first group of algorithms is represented by programs DOCK and FlexX, and the second group is represented by programs AutoDock and ICM. All these programs are publically available and were obtained by us under either academic (DOCK 4.0, AutoDock 3.0, and ICM 2.8) or demo (FlexX 1.8) licenses. In order to compare the performance of selected programs, we undertook two kinds of studies: cross–docking experiments involving protein–ligand complexes with different ligands but the same protein and database screening against the same proteins. We have chosen to perform the cross–docking experiments because they have the same underlying principle as the database screening experiments: in both cases different ligands are being fitted into

the same receptor. Alternatively, we could have used an uncomplexed proteins for docking experiments. However, this approach is problematic since the crystal structure of apoproteins is rarely available and in addition one would need to employ flexible receptor description. The dataset for docking experiments included 11 groups of complexes, each group containing between two and eight members. The library for screening experiments contains ligands from docking dataset and additional 10,000 molecules randomly selected from a database of commercially available compounds (ACD) distributed by MDL.

The outline of the paper is as follows. First we briefly review the algorithms employed by the programs. Next we describe input data preparation and docking protocols. Finally we present results on docking and screening experiments.

2 Algorithms

Here we briefly outline the docking algorithms employed by AutoDock, DOCK, FlexX, and ICM. The reader is referred to the original papers for a detailed account.

AutoDock

AutoDock explores the conformational space of the ligand using the Lamarkian genetic algorithm (LGA), which is a hybrid of a genetic algorithm (GA) with an adaptive local search (LS) method.[1] In this approach the ligand's state is represented as a chromosome, which is composed of a string of real-valued genes, describing ligand's location (three coordinates), orientation (four quaternions) and conformation (one value for each torsion). The simulation is started by creating a random population of individuals. It is followed by a specified number of generation cycles, each consisting of the following steps: mapping and fitness evaluation, selection, crossover, mutation and elitist selection. Each generation cycle is followed by a local search. The

solutions are scored using the energy scoring finction, which include terms accounting for short–ranged van der Waals and electrostatic interactions, loss of entropy upon ligand binding, hydrogen bonding and solvation.

DOCK and *FlexX*

Both DOCK[2] and FlexX[3,4] employ the incremental reconstruction algorithm. In this algorithm rigid anchor (DOCK) or base (FlexX) fragments are identified first. At the next step the selected fragment is placed into the active site of the receptor using sphere–matching procedure (DOCK) or hashing technique (FlexX). The complete ligand is constructed by adding the remaining components step by step. At each step of reconstruction a specified number of optimal partial solutions are selected for the next extension step. In DOCK the solutions are scored using energy, contact or chemical scoring functions. The energy scoring function, which was used in this study, includes van der Waals and electrostatic components. In FlexX the scoring is done using a modified Böhm scoring function, which includes the following terms: entropic, which accounts for loss of entropy upon ligand binding; hydrogen bonding; ionic, acounting for electrostatic interactions; aromatic, which accounts for interactions between aromatic groups; and lipophilic, which accounts for hydrophobic interactions. All terms, except the entropic, are scaled by the corresponding heuristic distance and angle dependent penalizing functions.

ICM

ICM performs flexible docking via Monte Carlo global optimization of the effective energy function in the internal coordinate space of the flexible ligand and flexible receptor.[5–7] The effective energy function includes the following terms: intramolecular van der Waals and torsion energy; modified intermolecular van der Waals interaction energy; hydrogen bonding term; term accounting for hydrophobic interactions; and electrostatic term using modified Coulomb law with distance dependent dielectric constant $\epsilon = 4r$.

The Monte Carlo procedure employed by ICM includes two types of moves: a pseudo–brownian positional move and a biased-probability multitorsion move. Each move is followed by full local energy minimization. Pseudo–Brownian random move changes the ligand position by moving the entire molecule with a certain amplitude and rotating it randomly around its center of mass by a certain angle. Biased-probability multitorsion move can be used to implement the flexible description of the receptor side–chains, however this capability was not invoked in the present study. Global optimization is performed starting from multiple starting points. The solutions are scored using the effective energy function.

3 METHODS

Input data preparation and algorithms comparison methodology

(1) The proteins which we have chosen for docking and database screening experiments, satisfy the following criterea: they have at least two entries with different ligands in the protein databank (PDB); they do not form covalent bonds with their respective ligands; majority of their ligands have relatively large number of rotatable bonds (see Table 1).

(2) The ligand input files were prepared according to the following precedure. First, we extracted ligand coordinates from the the corresponding PDB file and assigned chemical bonds, partial charges and added hydrogen atoms using ICM. All carboxylic acid and phosphoric acid groups were ionized and all amino–groups were protonated. Next, all torsion bonds were randomized and local minimization was performed. After that the ligand coordinates were modified in such a way that its center of geometry was superimposed with that of the reference ligand. Finally, the ligand coordinates were written into MOL2 and PDB format files.

(3) The receptor input files were prepared according to the following precedure. First, we removed from the corresponding PDB file all water molecules, ligand atoms and those ions, which did not belong to the active

site of the receptor. Next hydrogen atoms were added and partial charges were assigned using ICM. This was followed by a local minimization. Finally, the receptor coordinates were written into MOL2 and PDB format files.

(4) The docking experiments were performed on the same computer and CPU time required for docking was recorded. The docking protocols are described in the next section. We emphasize here, that all methods use the same receptor coordinates and start with exactly the same initial location, conformation and orientation of the ligands. The length of the docking experiments was controlled by the default or recommended parameter settings. We observed that doubling the length of the docking experiments does not improve significantly the accuracy of the solutions.

(5) For each docking method only the best scoring solution per complex was saved. Different algorithms were compared based on the root-mean square deviation (RMSD) of heavy atoms of the best predicted structures from the corresponding crystal structures. If the ligand has local topological symmetry at single bonds, whose torsion angle can be changed by a rotation of less than 360° without changes in the global conformation of the ligand, the RMSD of alternative orientations was calculated and the smallest one was kept for the purpose of comparison of different algorithms. The coordinates of ligand structure, used for RMSD calculations, were obtained by superimposing its the crystallographic coordinates protein coordinates with the receptor coordinates used for the docking.

(6) In order to quantify ligand docking quality and compare performance of different methods, we introduced the docking acciracy (DA) function, which makes use of RMSD values and measures how accurately the ligands – members of a particular group, – are docked by a given method:

$$DA = f_{rmsd\leq 2} + 0.5\,(f_{rmsd\leq 3} - f_{rmsd\leq 2})\,, \qquad (1)$$

where $f_{rmsd\leq a}$ indicates the fraction of ligands docked into a given receptor with RMSD less or equal a Å. The docking accuracy of the method for a

particular receptor is zero if $f_{rmsd \leq 3}$ is zero.

Docking protocols

In all algorithms studied here, the receptor is treated as a rigid body and a grid potential is used to evaluate the scoring functions. This simplification allows to perform docking more efficiently, which is especially crucial in database screening.

AutoDock

AutoDock requires the receptor and ligand coordinates in MOL2 format. Nonpolar hydrogen atoms were removed from the receptor file and their partial charges were added to the corresponding carbon atoms. The program *Mol2topdbqs* was used to transform receptor MOL2 file into PDBQS format file containing the receptor atoms coordinates, partial charges and solvation parameters. The program *AutoTors* was used to transform ligand MOL2 file into PDBQ file, merge nonpolar hydrogen atoms and define torsions. The grid calculations were setup with utility *Mkgpf3* and maps were calculated with program *AutoGrid.* The grid maps were centered on the ligand's binding site and they were $61 \times 61 \times 61$ points. The default parameters setting, generated by program *Mkdpf3*, was used for docking. For each complex 10 dockings were performed. The initial population was set to 50 individuals; maximum number of energy evaluations was 2.5×10^5; maximum number of generations was 27,000. The other parameters provided by the default setting were the same as in Ref. [1].

DOCK

DOCK requires the following receptor files: MOL2 format file containing coordinates of all atoms; PDB file containing heavy atoms coordinates only; and PDB file containing heavy atoms excluding those of the active site. The active site atoms included those receptor atoms which were within 6.5 Å from

the reference ligand atoms. The ligand coordinates were provided in MOL2 format. The site points for the ligand docking were identified using *SPHGEN* program. The number of docking points did not exceed 50. Energy score was employed for orientational and conformational search. Grid maps were calculated using *Grid* program, with grid spacing of 0.5 Å. The energy cutoff distance of 10 Å was employed. Electrostatic interaction were calculated with distance depending dielectric constant. The dielectric factor was set to 4. Proteins were represented by a united atom model. Flexible bonds and anchors were automatically identified by the DOCK. Conformational search was done with torsion drive. The clash overlap set to 0.5. Top 25 conformations were retained during each cycle of the search. Multiple anchors were allowed, with minimum number of heavy atoms in the anchor set to 10. Orientational search was performed with an automated matching. Maximum number of orientations was 500 for docking experiments and 100 for database search. The local energy minimization of orientations and conformations of ligand and anchor was performed. The ligand reminimization was turned on. The default minimization parameters were employed.

FlexX

FlexX requires MOL2 format file for the ligand and PDB format file for the receptor. The default settings as provided with FlexX 1.8 package were used for flexible docking and database screening. The conformational flexibility of the ligand is modeled by a discrete set of preferred torsional angles for acyclic single bonds. The rings were considered rigid, since the program *CORINA* for treating multiple conformations of the rings was not included in the distribution. The active site and the interaction surface of the receptor were defined by placing a reference ligand and using 6.5 Å cutoff distance. Base fragments were selected automatically. The maximum number of base fragments was 4. The base fragment was placed into the active site by using two algorithms. The first one superimposes triples of interaction centers of a base fragment with triples of compatible interactions

in the active site. The second algorithm, called matching, is used when the base fragment had fewer than three interaction centers. The maximum number of solutions retained for the next iteration step was 400.

ICM

Grid maps were calculated with a grid spacing of 0.5 Å. Docking was performed with a default script provided by ICM. During the docking, either one of the torsional angles of the ligand was randomly changed or pseudo–Brownian move was performed. Each random change was followed by 100 of local conjugate-gradient minimizaion. The new conformation was accepted or rejected according to Methropolis rule using temperature of 600 K. The length (number of Monte Carlo steps) of docking run as well as the length of local minimiztion length was determined automatically by the adaptive algorithm, depending on the size and number of flexible torsions in the ligand.

4 RESULTS

All docking experiments and database screening were performed on a SGI 10000 equipped with a single 195 MHz IP28 processor and 128 MB main memory.

Docking experiments

The following receptors were used for docking experiments: trypsin (PDB entry 3ptb), cytochrome P-450$_{cam}$ (PDB entry 1phf), neuraminidase (PDB entry 1nsc), carboxypeptidase A (PDB entry 1cbx), L-arabinose binding protein (PDB entry 1abe), ϵ-thrombin (PDB entry 1etr), thermolysin (PDB entry 3tmn), pencillopepsin (PDB entry 1apt), intestinal fat-acid binding protein (PDB entry 1icm), ribonuclease T_1 (PDB entry 1gsp), and carbonic anhydrase II (PDB entry 1cil). Most receptors have three ligand members with with exception of trypsin, which has 8 members and penicillopepsin, which has 2 members.

The complexes for which cross–docking experiments were performed and the docking results, such as RMSD values and CPU times, are given in Table 1. The docking accuracies of studied programs are summarized in Table 2. We see that ICM is the most accurate in predicting correct protein–ligand conformations. It gets perfect score of one for 5 receptors, which means that all members of those receptors are docked within RMSD less than 2 Å from the corresponding crystal structures. Other methods are much less accurate in their predictions. In particular, we observe that there are several receptors for wich they fail to produce any acceptable solution at all. Those are ϵ–thrombin, thermolysin and carbonic anhydrase II in case of AutoDock; neuraminidase, carboxipeptidase and pensillopepsin in case of DOCK; and thermolysine, pensillopepsin and carbonic anhydrase II in case of FlexX. On average the docking accuracy of AutoDock and FlexX is approximately the same and that of DOCK is slightly worse.

As evident from Table 2 the average docking time increases in the following order: FlexX, DOCK, ICM and AutoDock. The low docking speed of AutoDock suggests that at the present time it is not suitable for database screening on a single processor computer. We note, however, that compared to overall cost in the drug development process the computational cost is of lesser importance. Moreover, it is deemed to reduce owing to rapid developments in computer industy. Thus when comparing different docking algorithms more emphasis should be attached to their accuracy rather than the computational cost. With this in mind we conclude based on the results of docking experiments that the Internal Coordinates Method has the best docking performance among studied algorithms.

Before proceeding to the library screening results, we note that in all docking experiments we used receptor and ligand MOL2 files generated by ICM. However, according to AutoDock, DOCK and FlexX manuals, it is recommended to use SYBYL to generate the ligand and receptor (AutoDock and DOCK) input files. The only difference between input files generated by SYBYL and ICM is in partial atomic charges. ICM uses bond charge

increment method from MMFF94 (Halgren) to assign partial atomic charges. In order to be certain that our results are not influenced by partial charges provided by ICM, we repeated docking experiments for AutoDock, DOCK and FlexX using charges generated by SYBYL: Kollman charges for receptor atoms and Gasteiger charges for ligand atoms. Our findings indicate that the accuracy of AutoDock, DOCK and FlexX with SYBYL charges was the same as with ICM charges.

Database screening

The same proteins that were used for docking experiments have been chosen to perform screening of a ligand library. As was mentioned in introduction, this library contains ligands that were used for docking experiments and additional 10,000 molecules selected randomly from ACD library. The purpose of the screening experiments was to find out how well the programs would distinguish the original ligands of the complexes among all database molecules. This has a tremendous practical implications, since good docking algorithm allows to cut significantly the cost of drug discovery process by reducing the fraction of compounds of the ligand library that should be analyzed experimentally as potential drug candidates. In order to quantify the database screening we use the following virtual screening (VS) function:

$$VS = f_{\leq 1} + 0.5\,(f_{\leq 5} - f_{\leq 1})\,, \tag{2}$$

where $f_{\leq a}$ indicates the fraction of original ligands found within top a % of scanned solutions. The virtual screening function varies between 0 and 1. The value VS for a particular receptor molecule is 1 if all of its ligands are found within top 1 % of scanned solutions and 0 if no original ligands are found within top 5 % of scanned solutions. The results of screening performed by DOCK, FlexX and ICM are summarized in Table 3 and the VS values are given in table 4. First we note that out of 37 original ligands ICM places 19 ligands within top 1 % scanned solutions, while the corresponding

number for DOCK and FlexX is 8 and 11 respectively. The easiest receptor for screening experiments is L-Arabinose, since all algorithms assign very high scores (VS equals 1) to its original ligands. There are also receptors for which algorithms fail to place original ligands even within 5 % of top scoring solutions. There are 4 such receptors in case of DOCK, 3 receptors in case of FlexX and 2 receptors in case of ICM. The average value of the virtual screening function is 0.30, 0.32 and 0.6 for DOCK, FlexX and ICM respectively. This result suggest, that it is sufficient to consider top 5 % of best scoring solutions produced ICM as potential drug candidates, while in case of DOCK and FlexX more than 5 % of top scorers should be taken for experimental verification.

As a sidenote, we found that the database screening results with DOCK are improved somewhat if the chemical scoring function is employed instead of energy score. This contrasts the results of docking experiments, where no influence of the the choice of scoring function was found.

5 CONCLUSIONS

Our results indicate that ICM outperforms other docking programs in both docking and database screening experiments.

References

[1] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson *J. Comp. Chem.*, **19**, 1639, (1998).

[2] S. Makino and I. D. Kuntz, *J. Comp. Chem.*, **18**, 1812, (1997).

[3] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, *J. Mol. Biol.*, **261**, 470, (1996).

[4] B. Kramer, M. Rarey, and T. Lengauer, *Proteins*, **37**, 228, (1999).

[5] R. Abagyan, M. Totrov, and D. Kuznetsov, *J. Comp. Chem.*, **15**, 488, (1994).

[6] M. Totrov, and R. Abagyan, *Proteins*, **Suppl. 1**, 215, (1997).

[7] M. Totrov, and R. Abagyan, in *The Thermodynamics of Protein-Ligand Interactions*; ???., Eds.; Wiley and Sons: New York, 2000 (in press).

Table 1 Results of the Docking Experiments

Ligand	Nrot[a]	AutoDock[b]	DOCK[b]	FlexX[b]	ICM[b]
Trypsin					
3ptb	3	0.80 (323)	0.59 (23)	1.11 (15)	0.49 (77)
1tng	2	0.62 (270)	0.86 (20)	1.08 (5)	0.71 (79)
1tnj	3	1.21 (290)	1.56 (22)	1.73 (35)	2.17 (21)
1tnk	4	1.69 (330)	1.87 (29)	1.70 (11)	2.53 (26)
1tni	5	2.61 (367)	5.26 (35)	2.73 (12)	3.40 (39)
1tnl	1	0.41 (300)	2.08 (25)	3.74 (30)	1.93 (17)
1tpp	7	1.80 (330)	3.25 (46)	1.95 (47)	1.71 (53)
1pph	11	5.14 (920)	3.91 (212)	3.27 (51)	1.44 (207)
Cytochrome P-450$_{cam}$					
1phf	1	2.09 (254)	2.39 (25)	4.68 (72)	1.23 (28)
1phg	5	3.52 (390)	5.57 (39)	4.87 (169)	0.46 (57)
2cpp	3	3.40 (230)	2.48 (22)	0.44 (6)	2.53 (34)
Neuraminidase					
1nsc	12	1.40 (610)	4.86 (97)	6.00 (57)	1.80 (119)
1nsd	11	1.20 (600)	4.51 (110)	1.56 (88)	1.04 (70)
1nnb	11	0.92 (650)	4.51 (88)	0.92 (71)	1.09 (108)
Carboxypeptidase A					
1cbx	5	1.33 (354)	3.13 (45)	1.32 (24)	0.82 (40)
3cpa	8	2.26 (512)	6.48 (56)	1.51 (172)	0.77 (51)
6cpa	16	8.30 (1007)	8.30 (163)	9.83 (81)	1.60 (350)
L-Arabinose binding protein					
1abe	4	0.16 (340)	1.87 (32)	0.55 (30)	0.36 (38)
1abf	5	0.48 (320)	3.25 (36)	0.76 (35)	0.61 (37)
5abp	6	0.48 (400)	3.89 (43)	4.68 (29)	0.88 (42)
ϵ-Thrombin					
1etr	15	4.61 (1153)	6.66 (371)	7.26 (104)	0.87 (444)
1ets	13	5.06 (1366)	3.93 (522)	2.11 (69)	6.22 (344)
1ett	11	8.12 (1003)	1.33 (371)	6.24 (72)	0.99 (219)

Continuation of **Table 1.**

Thermolysin					
3tmn	10	4.51 (630)	7.09 (107)	5.30 (67)	1.36 (99)
5tln	14	5.34 (711)	1.39 (140)	6.33 (62)	1.42 (196)
6tmn	20	8.72 (1027)	7.78 (262)	4.51 (67)	2.60 (420)
Penicillopepsin					
1apt	30	1.89 (1242)	8.06 (416)	5.95 (76)	0.88 (700)
1apu	29	9.10 (1002)	7.58 (409)	8.43 (78)	2.02 (590)
Intestinal FABP					
1icm	13	1.80 (583)	3.99 (112)	2.94 (31)	1.11 (154)
1icn	17	3.99 (583)	3.88 (166)	2.95 (42)	1.35 (314)
2ifb	15	3.09 (513)	1.43 (135)	8.94 (14)	1.04 (234)
Ribonuclease T_1					
1gsp	4	2.67 (592)	1.16 (59)	3.71 (44)	0.54 (59)
1rhl	7	0.96 (710)	0.71 (72)	1.15 (51)	3.53 (78)
1rls	7	0.98 (703)	1.75 (76)	4.33 (72)	0.79 (77)
Carbonic Anhydrase II					
1cil	6	5.81 (460)	2.78 (63)	3.52 (87)	2.00 (58)
1okl	5	8.54 (396)	5.65 (38)	4.22 (105)	3.03 (42)
1cnx	13	10.9 (700)	7.35 (63)	6.83 (72)	2.09 (176)

a) Number of rotatable bonds in the ligand.

b) First number is RMSD values in Å; the second number (in parentheses) is docking time in seconds.

Table 2 Docking accuracies and average docking times (seconds) of algorithms

Receptor	AutoDock	DOCK	FlexX	ICM
Trypsine	0.81 (391.2)	0.57 (51.5)	0.73 (25.7)	0.75 (64.9)
Cytochrome $P450_{cam}$	0.17 (291.3)	0.33 (28.7)	0.33 (82.3)	0.83 (39.7)
Neuraminidase	0.67 (620)	0 (98.3)	0.67 (72)	1.00 (99)
Carboxypeptidase	0.50 (624.3)	0 (88)	0.67 (92.3)	1.00 (147)
L-Arabinose	1.00 (353.3)	0.33 (37)	0.67 (31.3)	1.00 (39)
ϵ–Trombin	0 (1174)	0.33 (421.3)	0.17 (81.7)	0.67 (335.7)
Thermolysin	0 (789.3)	0.33 (169.7)	0 (65.3)	0.83 (238.3)
Pencillopepsin	0.50 (1122)	0 (412.5)	0 (77)	1.00 (645)
Intestinal Fat-Acid	0.33 (559.7)	0.33 (137.7)	0.33 (29)	1.00 (234)
Ribonuclease T_1	0.83 (668.3)	1.00 (69)	0.33 (55.7)	0.67 (71.3)
Carbonic Anhydrase II	0 (518.7)	0.17 (54.7)	0 (88)	0.67 (92)
Average	0.44 (646.5)	0.31 (142.6)	0.39 (63.7)	0.93 (182.3)

Table 3 Results of the Virtual Database Screening

Ligand	DOCK[a]	FlexX[a]	ICM[a]
Trypsin		9928	
3ptb	2014 (20.3)	1391 (14)	15 (0.1)
1tng	2855 (28.7)	3056 (31)	27 (0.3)
1tnj	4478 (45.1)	4757 (48)	106 (1.1)
1tnk	5771 (58.1)	3685 (37)	1463 (14.6)
1tni	4330 (43.6)	3821 (38)	60 (0.6)
1tnl	8138 (81.9)	3880 (39)	5181 (51.6)
1tpp	393 (4.0)	27 (0.3)	45 (0.4)
1pph	37 (0.4)	63 (0.6)	32 (0.3)
Cytochrome P-450_{cam}		9928	
1phf	2676 (26.9)	1309 (13.2)	1010 (10.1)
1phg	6022 (60.6)	141 (1.4)	1280 (12.7)
2cpp	724 (7.3)	1250 (12.6)	1171 (11.7)
Neuraminidase		8418	
1nsc	1893 (19.0)	611 (7.2)	503 (5.01)
1nsd	933 (9.4)	252 (3.0)	44 (0.44)
1nnb	664 (6.7)	116 (1.4)	43 (0.43)
Carboxypeptidase A		8951	
1cbx	3656 (36.8)	240 (2.7)	23 (0.2)
3cpa	1411 (14.2)	122 (1.4)	32 (0.3)
6cpa	842 (8.5)	246 (2.7)	360 (3.4)
L-Arabinose binding protein		7593	
1abe	52 (0.5)	19 (0.2)	1 (0.01)
1abf	119 (1.2)	15 (0.2)	3 (0.03)
5abp	21 (0.2)	29 (0.4)	2 (0.02)
ϵ-Thrombin		5579	
1etr	556 (5.6)	23 (0.4)	157 (1.6)
1ets	38 (0.4)	13 (0.2)	672 (6.7)
1ett	90 (0.9)	485 (8.7)	6 (0.06)

Continuation of **Table 3.**

Thermolysin		8353	
3tmn	3412 (34.3)	289 (3.5)	
5tln	6554 (66.0)	53 (0.6)	
6tmn	1494 (15.0)	707 (8.5)	
Penicillopepsin		8778	
1apt	28 (0.3)	3637 (41)	110 (1.1)
1apu	377 (3.8)	7036 (80)	1286 (12.8)
Intestinal FABP		5903	
1icm	1008 (10.1)	3562 (70.8)	266 (2.6)
1icn	335 (3.4)	3856 (75.1)	48 (0.5)
2ifb	704 (7.1)	5097 (99.3)	49 (0.5)
Ribonuclease T_1		9870	
1gsp	442 (4.4)	1060 (10.7)	59 (0.6)
1rhl	553 (5.6)	1009 (10.2)	45 (0.4)
1rls	6849 (68.9)	735 (7.4)	9090 (90.6)
Carbonic Anhydrase II		9857	
1cil	3476 (35.0)	364 (3.7)	7193 (71.7)
1okl	1841 (18.5)	1907 (19.3)	3105 (30.9)
1cnx	29 (0.3)	1429 (14.5)	9090 (90.6)

a) First number is a rank of the original ligand of the receptor; second number (in paretheses) is a corresponding fraction in the total database.

Table 4 Virtual screening functions

Receptor	DOCK	FlexX	ICM
Trypsine	0.19	0.25	0.75
Cytochrome $P450_{cam}$	0	0.33	0
Neuraminidase	0	0.33	0.83
Carboxypeptidase	0	0.33	0.83
L-Arabinose	1	1	1
ϵ–Trombin	0.66	0.66	0.5
Thermolysin	0	0.5	
Pencillopepsin	0.75	0	0.5
Intestinal Fat-Acid	0.17	0	0.83
Ribonuclease T_1	0.17	0	0.67
Carbonic Anhydrase II	0.33	0.17	0
Average	0.30	0.32	0.6